NEAR EAST UNIVERSITY

GRADUATE SCHOOL OF HEALTH SCIENCES

BIOSTATISTICS DEPARTMENT

BAYESIAN MACHINE LEARNING ANALYSIS WITH MARKOV CHAIN MONTE CARLO TECHNIQUES FOR ASSESSING CHARACTERISTICS AND RISK FACTORS OF COVID-19 IN ERBIL CITY-IRAQ 2020-2021

HEWIR KHIDIR

Ph.D. THESIS

NICOSIA

(2023)

HEWIR KHIDIR

NEAR EAST UNIVERSITY GRADUATE SCHOOL OF HEALTH SCIENCES

FACULTY OF MEDICINE

DEPARMENT OF BIOSTATISTICS

Ph.D. THESIS

THESIS SUPERVISOR

PROF. DR. ILKER ETIKAN

NICOSIA

(2023)

Acceptance/Approval

We as the jury members certify the "Bayesian Machine Learning Analysis with Markov Chain Monte Carlo Techniques for assessing characteristics and risk factors of Covid-19 in Erbil City-Iraq 2020-2021". "Prepared by Hewir Khidir on ....has been found satisfactory for the award of Degree of PhD.

Thesis Committee;

Chair of the committee: **Prof. Dr. İlker ETİKAN (Advisor)**
Near East University
Sig: ....

Member: **Prof. Dr. Beyza ŞAHİN**
Pamukkale University
Sig: ....

Member: **Prof. Dr. Selim Yavuz SANİSOĞLU**
Ankara Yıldırım Beyazıt University
Sig: ....

Member: **Assoc. Prof. Dr. Özgür TOSUN**
Near East University
Sig: ....

Member: **Assoc. Prof. Dr. Uğur BİLGE**
Akdeniz University
Sig: .... Bilge

Approved by: **Prof. Dr. K. Hüsnü Can BAŞER**
Director of Institute of Graduate Studies
Near East University
Sig: ....

Declaration

I solemnly affirm that every piece of information contained within this document adheres rigorously to the principles of academic integrity and ethical conduct. Furthermore, in strict compliance with these standards, I have meticulously attributed and referenced all non-original content and findings as mandated.

Date:

Signature:

Name, Surname: Hewir Khidir

Acknowledgments

First of all, I would like to thank my supervisor Prof. Dr.İlker Etikan (Near East University) for his supervision, support, and sharing his knowledge with me during my thesis work.

I would like to thank, Prof. Dr. Yavuz Sanisoğlu (Yıldırım Beyazıt Unversity) and Assist. Prof. Dr. Özgür Tosun (Near East University), for their support and motivations.

Dedication

To my unwavering source of inspiration, my family, whose boundless love, encouragement, and support have illuminated my path throughout this academic journey. Your unwavering faith in me has been the driving force behind the completion of this thesis. This accomplishment is as much yours as it is mine.

Abstract

Bayesian Machine Learning Analysis with Markov Chain Monte Carlo Techniques for Assessing Characteristics and Risk Factors of Covid-19 in Erbil City-Iraq 2020-2021

Hewir Khidir

Ph.D., Department of Biostatistics

June, 2023, 128 pages

Study's Background

The study aims to showcase machine learning techniques in the application of medical datasets for improving identification of correlations and relationships between variables, which will lead to more informed decision-making. Unlike other studies, intensive statistical modelling is used to understand and find the effective of variables cause to lead death due to Covid-19. Due to large dataset, not common approaches derive us to ideal conclusion. Furthermore, Bayesian technique is applied to generate predictive posterior distributions of the unknown parameters in the model in neural network as well as logistic regression, which helps us to avoid overfitting in machine learning applications and have additional measurements in assessing fitted model performance. The study primarily focuses on analysing the patient profile of those who were hospitalized due to Covid-19 infection. To achieve this, two hospitals with a significant number of Covid-19 cases since the beginning of the pandemic have been chosen for analysis - one from the public sector and the other from the private sector.

Methods

The pattern and distribution of the study variables will be defined using descriptive statistics. The basic statistical measurements are calculated such as, mean, median and standard deviation of quantitative variables, whereas the frequencies and percentages are used to decide the trend of categorical variables. The dataset comprises an extensive array of variables, a few of which are outlined as age, gender, Smoking, Fever, Cough, Sputum, Hypertension, Diabetes, Stroke, Temperature, HR, Respiratory, SpO2, Quadrant, Pulmonary.

WBC Count, Neutrophil, Lymphocyte, Platelet, Albumin, Creatinine, CRP, APTHT, Fibrinogen, D dimer.

Results

According to the results extracted from the statistical analysis, the Bayesian neural network demonstrated superior performance in terms of classification measurements such as AUC (84.66%), F1 (87.11%), Precision (88.29%), and Recall (85.96%). The Bayesian logistic regression also performed well, but with slightly lower scores, achieving AUC (83.07%), F1 (85.59%), Precision (84.55%), and Recall (85.59%). In contrast, kNN algorithm had the worst performance with very low scores (AUC=52.38%, F1=57.55%, Precision=57.01%, Recall=58.10%). Regarding the variables' association with mortality, stepwise forward selection helped to identify seven significant variables. Age was found to be the most significant variable in predicting the probability of dying, with patients in the age group of (18-44) having 12 times higher odds, patients in the age group of (45-64) having 123 more odds, and patients above 65 years old having 436 times more chance to die compared to patients below 18 years old. Severe coughing was also significant with 7.26 odds, and patients suffering from diabetes had 2.82 times more chance to die. Moreover, SpO2 contributed to a decrease of 20% in the relative risk of dying from Covid-19 disease. Gender and Smoking did not show a significant association with mortality.

Finally, the Bayesian approach showed higher sensitivity and specificity than the classic neural network.

Keywords:

Covid-19, Bayesian Neural Network, Bayesian Logistic Regression Modelling, Markov Chain Monte Carlo Odds Ratio, ROC Curve, Classification Measurements,

Özet

Erbil Şehri-Irak'ta Kovid-19'un Özelliklerini ve Risk Faktörlerini Değerlendirmek için Markov Zinciri Monte Carlo Teknikleri ile Bayesian Makine Öğrenimi Analizi 2020-2021

Doktora, Biyoistatistik Anabilim Dalı

Haziran, 2023, 128 sayfa

Çalışmanın Arka Planı

Çalışma, değişkenler arasındaki korelasyonların ve ilişkilerin tanımlanmasını iyileştirmek için tıbbi veri kümelerinin uygulanmasında makine öğrenimi tekniklerini göstermeyi amaçlıyor ve bu da daha bilinçli karar vermeye yol açacak. Diğer çalışmalardan farklı olarak Kovid-19 nedeniyle ölüme yol açan değişkenlerin anlaşılması ve etkisinin bulunması için yoğun istatistiksel modelleme kullanılıyor. Büyük veri seti nedeniyle ortak yaklaşımlar bizi ideal sonuca ulaştırmamaktadır. Ayrıca, sinir ağındaki modeldeki bilinmeyen parametrelerin tahmine dayalı sonsal dağılımlarını oluşturmak için Bayesian tekniği uygulanıyor ve lojistik regresyon, makine öğrenimi uygulamalarında aşırı uyumdan kaçınmamıza ve uygun model performansını değerlendirmede ek ölçümlere sahip olmamıza yardımcı oluyor. Çalışmada öncelikle Kovid-19 enfeksiyonu nedeniyle hastaneye kaldırılanların hasta profilinin analiz edilmesine odaklanılıyor. Bunu başarmak için, salgının başlangıcından bu yana önemli sayıda Kovid-19 vakasının görüldüğü, biri kamu sektöründen, diğeri özel sektörden olmak üzere iki hastane analiz için seçildi.

Yöntemler

Çalışma değişkenlerinin şekli ve dağılımı tanımlayıcı istatistikler kullanılarak tanımlanacaktır. Niceliksel değişkenlerin ortalama, ortanca ve standart sapması gibi temel istatistiksel ölçümler hesaplanırken, kategorik değişkenlerin eğilimini belirlemek için frekans ve yüzdeler kullanılır. Veri seti, birkaçı yaş, cinsiyet, Sigara içme, Ateş, Öksürük, Balgam, Hipertansiyon, Diyabet, İnme, Sıcaklık, HR, Solunum, SpO2, Quadrant, Pulmoner olarak özetlenen çok çeşitli değişkenlerden oluşur. WBC Sayısı, Nötrofil, Lenfosit, Trombosit, Albümin, Kreatinin, CRP, APTHT, Fibrinojen, D dimer.

Sonuçlar

İstatistiksel analizden elde edilen sonuçlara göre Bayes sinir ağı, AUC (%84,66), F1 (%87,11), Precision (%88,29) ve Recall (%85,96) gibi sınıflandırma ölçümleri açısından üstün performans sergiledi. Bayesian lojistik regresyonu da iyi performans gösterdi ancak biraz daha düşük puanlarla AUC (%83,07), F1 (%85,59), Precision (%84,55) ve Recall (%85,59) elde etti. Buna karşılık kNN algoritması çok düşük puanlarla en kötü performansı gösterdi (AUC=%52,38, F1=%57,55, Precision=%57,01, Recall=%58,10). Değişkenlerin mortalite ile ilişkisine ilişkin olarak, adım adım ileri seçim, yedi anlamlı değişkenin belirlenmesine yardımcı oldu. Yaşın, ölüm olasılığını tahmin etmede en anlamlı değişken olduğu, yaş grubundaki hastaların (18-44) oranlarda 12 kat, (45-64) yaş grubundaki hastalarda ise 123 kat daha fazla oran elde edildiği, ve 65 yaş üstü hastaların ölme şansı, 18 yaş altı hastalara göre 436 kat daha fazladır. Şiddetli öksürük de 7,26 oranla anlamlıydı ve diyabetli hastaların ölme şansı 2,82 kat daha fazlaydı. Üstelik SpO2, Kovid-19 hastalığından göreceli ölüm riskinde %20'lik bir azalmaya katkıda bulundu. Cinsiyet ve Sigara kullanımı mortaliteyle anlamlı bir ilişki göstermedi.

Son olarak Bayes yaklaşımı, klasik sinir ağından daha yüksek hassasiyet ve özgüllük gösterdi.

Anahtar Kelimeler:

Covid-19, Bayesian Sinir Ağı, Bayesian Lojistik Regresyon Modellemesi, Markov Zinciri Monte Carlo Odds Oranı, ROC Eğrisi, Sınıflandırma Ölçümleri,

Table of Contents

Acceptance/Approval 3
Declaration 4
Acknowledgments 5
Dedication 6
Abstract 7
    Study’s Background 7
    Methods 7
    Results 8
Özet 9
List of Abbreviations 16
CHAPTER I 18
INTRODUCTION 18
    Background Of the Study 24
    Purpose of Study 25
    Research Questions 26
    Significance of Study 26
Limitation 27
LITERATURE REVIEW 28
    Introduction 28
    The Significance of the Occurrence and Incidence of Covid-19 28
    The Level of Risk Factors of Covid-19 29
    Covid-19's Impact on Bodily Organs 31
    Machine Learning Algorithms, and Traditional Statistical Analysis Methods 35
    MCMC in Neural Network Modelling 37
    Monte Carlo Simulation for Epidemics 38
    Neural Networks for Forecasting 39
    Covid-19 Model Complexities 40

CHAPTER III 41

Methodology and Statistical Learning of Bayesian Inference 41

- Methodology 41
- Research Method 41
- Data Capture and Analysis Strategies 42
- Variables Description 43
- Statistical Methods and Techniques 44
- Validity 44
- Generalizability 44
- Odds Ratio: 44
- Confusion matrix. 45
- Receiver Operating Characteristic Curve 47
- The Bayesian Logistic Regression Model 47
- Artificial Neural Network Analysis 56
- Bayesian Neural Network 62
- Other Machine Learning Methods 69
- Summary 72

CHAPTER IV 73

Statistical Analysis Results 73

- Descriptive and Visualization Analysis 73
- Response Association with Quantitative Variable 73
- Response Association with Qualitative Variables 76
- Model Building Analysis: 81
- MCMC Result Assessment (Posterior Distribution) 91
- Model Accuracy and Diagnosis Assessment 93
- Implementing Other Machine Learning Methods 96

  Result of Bayesian Neural Network and Classic Neural Network Analysis 97
  Relative Importance Analysis of Bayesian Neural Network 99
  Confusion Matrix Analysis of Applied Models 101
  Summary 103
CHAPTER V 104
Discussion And Analysis 104
CHAPTER VI 108
Conclusion And Recommendation 108
  Conclusion 108
  Recommendation 109
REFERENCES 110
Appendix A: 121
  Full package of Bayes by Backprop Algorithm in Bayesian Neural Network Application 121
Appendix B: 122
  Stepwise Forward Selection 122
Appendix C: 124
  Results of Fitted Models 124
Appendix D: 126
  Density illustration of the non- significant variables 126
Appendix E: 127
  Bar chart of non-significant categorical variables 127
Appendix F 128

List of Tables

Table 1: Variable Descriptions of the Dataset........43
Table 2: Confusion Matrix Exploration........45
Table 3: Descriptive Statistics of Laboratory Results Under Investigation Recovered and Died Cases........74
Table 4: Descriptive Statistics of Categorical Variables Associated with Response Variable........78
Table 5: Univariate Logistic Regression Analysis........83
Table 6: Diagnosis Test Result For Univariate Logistic Regression Models (MLE And MCMC) Parameter Estimation........84
Table 7: Accuracy Analysis for Univariate Logistic Regression models (MLE and MCMC) parameter estimation........87
Table 8: Result of stepwise forward model selection approach for Bayesian logistic regression (MCMC) and classic logistic regression (MLE)........88
Table 9: The best fitted model outcomes by Bayesian logistic regression (MCMC parameter estimation)........89
Table 10: Odds Ration and 95% Confidence Interval of Odds Ratio for Bayesian logistic regression coefficients estimated by MCMC........90
Table 11: Credible Interval Finding of Posterior Parameters........91
Table 12: Descriptive Statistics (Minimum, Maximum, Mean, Median, SD) of the posterior findings........93
Table 13: Test of Goodness-of-fit for the final model........93
Table 14: Confusion Matrix Result of Bayesian Logistic Regression........94
Table 15: Model Performance Analysis of Applied Algorithms........97
Table 16: Confusion Matrix Result of Bayesian Neural Network and Classic Neural Network........97
Table 17: Posterior Weight Matrix of Bayesian Neural Network........100
Table 18: Result of Relative Importance Measurement........100
Table 19: Confusion Matrix Outputs of Computed Models........102

List of Figures

Figure 1: Human’s Body with description of Covid19 Impact 32

Figure 2: Feed-forward neural network illustration with 2 hidden layers 58

Figure 3: ReLU Activation Function Demonstration 59

Figure 4: SIGMOID Activation Function Demonstration.. 60

Figure 5: Left: Bayesian Neural Network with probability distribution over weights. Right: Classic Neural Network with point estimates for weights. 62

Figure 6: Box-Plot Illustration for continuous covariates 75

Figure 7: Density Distribution with respect of response variable 77

Figure 8: Bar chart presentation of Age with respect to Response variable levels 79

Figure 9: Cough vs Response Bar plot B) Fever vs Response bar plot 80

Figure 10: Bar charts for A) Diabetes, B) Hypertension, C) Sputum 81

Figure 11: Correlation Heatmap Matrix For All Variables Study.. 82

Figure 12: Association of Quantitative variables with response using Logistic Regression (MCMC) 85

Figure 13: ROC Evaluation Curves for applied univariate models in Bayesian logistic regression 86

Figure 14: Trace plots and Histogram charts of posterior distribution estimated for Bayesian logistic regression model (MCMC) 92

Figure 15: ROC Curve characteristics for Out-sample and In-sample dataset run in Bayesian Logistic Regression Model 95

Figure 16: Bayesian Logistic Regression Assessment Graph 95

Figure 17: Bayesian Neural Network Architecture for The Study 98

Figure 18: ROC plot of Bayesian Neural Network and Classic Neural Network 99

Figure 19: Relative Importance of Inputs with Analyzing Weight Matrix 101

Figure 20: ROC Evaluation Curve of all eight models applied to predict risk of dying from Covid-19 103

List of Abbreviations

Terms	Definition
Covid-19	Corona Virus Disease 2019
ML:	Machine Learning
OR:	Odds Ratio
CI:	Confidence Interval
ROC:	Receiver Operating Characteristic
LR (MLE):	Classic Logistic Regression
LR (MCMC):	Bayesian Logistic Regression
BANN:	Bayesian Artificial Neural Network
ANN:	Artificial Neural Network
WHO:	World Health Organization
CR:	Creatine
WBC:	White Blood Cell
PLT:	Platelet
SpO2	Oxygen
DIC:	Disseminated Intravascular Coagulation
FDPs	Fibrin Degradation Products
CAD	Coronary Artery Disease
TMPRSS2	Transmembrane Serine Protease 2
ACE2	Angiotensin-Converting Enzyme 2
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings
HMC	Hamiltonian Monte Carlo
KL-divergence	Kullback-Leibler Divergence
ELBO	Evidence Lower Bound
SVI	Stochastic Variational Inference
AUC	Area Under ROC Curve
ROC	Receiver Operating Characteristic Curve
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SEIR	Susceptible-Exposed-Infectious-Recovered

ABC	Accurate Bayesian Computation
SDGs	Sustainable Development Goals
KRI	Kurdistan Region Of Iraq
NN	Neural Network
NGOs	Non-Governmental Organizations
EHRs	Electronic Health Records
IL-6	Interleukin-6
FDPs	Fibrin Degradation Products
DIC	Disseminated Intravascular Coagulation
ICU	Intensive Care Unit
BCNN	Bayesian Convolutional Neural Networks
RF	Random Forest
ET	Extreme Randomization Of Trees
MC	Monte Carlo
EMS	Emergency Medical Services
GROOMS	Group Optimized And Multisource Selection
CMC	Composite Monte-Carlo
BGFS-PNN	Bayesian Generalized Fast Stochastic Pnn
LSTMs	Long Short-Term Memory
SD	Standard Deviation
TP	True Positives
FP	False Positives
TN	True Negatives
FN	False Negatives

CHAPTER I

INTRODUCTION

In December 2019, a newly emerged infectious disease called Coronavirus disease 2019 (Covid-19), which was caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was first identified in Wuhan, China. In Iraq, Covid-19 was first detected in Southern Iraq in February 2020 and as a result of the emergence of these incidents, the Kurdistan Region enacted stringent security measures. To determine the limit of spreading the virus in the region, several indicators were used such as, the closure of schools and colleges, closure of borders and airports, cancellation of civic and religious events, and obligatory quarantine for people returning from trips abroad and encounters. More than 1.2 million people have died because of Covid-19, a new coronavirus disease that has affected over 45 million individuals globally (Organization, 2020a).

A broad spectrum of symptoms can be caused by Covid-19, as 14 percent and 5 percent of the patients whose test results were confirmed had Covid-19 cases that were either severe or serious. The emerge damaged the healthcare system because it spread very quickly and was the main cause of death, consequently most of the available medical resources were secured to contain the virus.

Some of the common symptoms identified initially were dyspnea, breathing rates below 30 breaths per minute, blood oxygen saturation levels under 93%, partial pressure of arterial oxygen to fraction of inspired oxygen ratio of less than 300 mmHg, and/or lung infiltrates of over 50%. These indicators typically appeared within 24 to 48 hours (Guo et al., 2020). Patients infected with severe respiratory issues required mechanical assistance for breathing and needed to be transferred to the critical care unit due to conditions such as shock, disseminated coagulopathy, or multiple organ failures.

Numerous risk factors were considered to be effective including age, gender, ethnicity, as well as nutrition, lifestyle, and laboratory indicators. It was generally agreed that the risk factors could help identify individuals who were more likely to contract the virus severely and face a higher risk of mortality. However, it was crucial to acknowledge that certain studies examine general risk factors associated with disease progression, whereas others focus on specific risk factors contributing to the advancement of Covid-19 into a critical stage (N. Chen et al., 2020).

In the realm of artificial neural networks (ANNs), a sophisticated connection system is established, which emulates human learning patterns to predict and comprehend future data. The groundwork for ANNs was laid in 1943 by Warren McCullock and Walter Pits, who introduced their initial model (Guan et al., 2020). Since then, ANNs have found success in various domains, surpassing human capabilities in prediction tasks such as decision support for cancer, streamflow forecasting, and weather prediction.

The fundamental building blocks of an ANN are neurons, interconnected in a manner that allows signals to flow between them and to other neurons. The collaboration of numerous neurons in an ANN facilitates its predictive capabilities, wherein layers play a crucial role. A layer comprises one or more neurons or groups of neurons, and ANNs can have an infinite number of such layers. To enhance prediction accuracy, ANNs often incorporate multiple hidden layers between their input and output layers (Liu et al., 2020). These hidden layers effectively transform incoming data into a format suitable for processing by the output layer.

At the core of the ANN, the input layer receives data in the form of real numbers. During the training process, the connection weights between neurons are iteratively adjusted to optimize performance and achieve the best possible outcomes.

After receiving feedback from the interconnected neurons in the underlying layers, an eager nerve cell employs an activation function to generate an output value for that specific layer. The true potential of an Artificial Neural Network (ANN) lies in its ability to accelerate the neuron firing process, enhancing its performance significantly (Huang et al., 2020). The selection of activation functions is often based on the problem at hand or determined through rigorous testing to identify the most suitable one.

Training various neural network models demands distinct approaches. Consequently, it is imperative to examine the key characteristics and patterns of major diseases' propagation to guide effective public health interventions and forecast the spread of potential new epidemics. In such scenarios, systems of ordinary differential equations prove invaluable, enabling the tracking of different population segments during the course of an infectious disease outbreak (G. Chen et al., 2020; N. Chen et al., 2020; Liu et al., 2020).

In the case of new infectious diseases like Covid-19, many of their crucial characteristics remain unknown and necessitate thorough investigation before accurate projections can be made. Having the lack of sufficient reliable data at the initial stages of an epidemic, it becomes challenging to identify these traits at the outset. Consequently, to obtain quick estimates during the first wave of the Covid-19 outbreak, model-based inference was employed, as extracting key epidemiological parameters directly from primary clinical monitoring data posed difficulties. One widely used approach involved SEIR (Susceptible-Exposed-Infectious-Recovered) models, which enabled predictions of the number of reported cases outside of Wuhan and the time it would take for the epidemic to double in size (J. T. Wu et al., 2020). These methods were also utilized to assess Covid-19's potential impact on children in various scenarios. The primary objective of such studies, along with others, was to identify critical epidemiological factors, including age-specific death rates, the impact of disease control measures on transmission, and the presence of unreported cases (Liu et al., 2020).

The application of models of the SIR type is an essential consideration to make. In order to forecast how an epidemic will develop in response to changes in population behaviour or climatic conditions, these models make assumptions about significant epidemiological traits and use trustworthy methods for quantifying uncertainty. In the fields of medical decision modelling and health policy, the process of estimating hidden model parameters by looking at model outputs is also known by the term "model calibration" (inverse inference) (Kompa et al., 2021). The conventional understanding of model calibration is that determining the optimal values for the parameters is an optimization problem (e.g., by performing non-linear least squares minimization or relying on maximum likelihood estimation). On the other hand, optimum and maximum likelihood approaches have a tendency to concentrate on point estimates for each parameter, but they lack the appropriate instruments to ensure that these point estimates are accurate. This is a significant challenge since reliable measures of uncertainty are necessary before making predictions about the future.

Therefore, accurate predictions had became point of concern in order to contain the disease and find proper medicine for patients. Bayesian statistical methods plays crucial roles and alongside of classical approach for model fitting, and several studies with machine learning algorithms have been studied to investigate Covid-19 risk factors as well as the trend of uprising the spread by (Abdulkareem et al., 2022; Dinar et al., 2022; Hameed Abdulkareem et al., 2022; Obaid et al., 2020; Saeed et al., 2022). This is because rather than providing a single-point estimate for the unknown parameters, Bayesian methods also offer the whole posterior distributions for the unknown parameters and Markov chain Monte Carlo sampling is one of the prevalent approaches to calibrating Bayesian models. For instance, gaussian process regression can be applied to estimate parameters for predicting temperature of doped Fe-based superconductors based on structural and topological parameters as studied by (Zhang & Xu, 2020, 2021b) and also the approach was compared with artificial neural network as studied by (Zhang & Xu, 2021a, 2021c). Moreover, Bayesian calibration is a more accurate method, and due to the complexity of the likelihood function, it is difficult to calculate the model's parameters straightforwardly (Khudhur & Kadir, 2022; Liu et al., 2021). Accurate Bayesian computation (ABC) can be applied to estimate the posterior distribution of unknown parameters when the likelihood function is hard to calculate or not known. ABC approaches are known for being inaccurate at dealing with large amounts of data and complicated models. This means that they can only be used for very simple models with a limited number of parameters.

Because of an outbreak of the novel coronavirus, the World Health Organization declared a global pandemic on March 11, 2020. The term was then abbreviated as Covid-19, or coronavirus disease 2019, and was used to refer to the ailment in subsequent years (Organization, 2020b). The first situation report on the sickness from the World Health Organization (WHO) was published in. Four countries, including China (278 instances), Thailand (2), Japan (1), and the Republic of Korea (1) (Organization, 2020a), have reported a total of 282 confirmed cases of the 2019-nCoV as of January 20th, 2020. (Organization, 2020a). They also were contributing their thoughts on crucial Covid-19-related issues. The number of cases, the number of fatalities, surveillance, strategic response goals, preparedness and reaction, as well as suggestions and guidance for the general public, were some of these subjects. Weekly operational and epidemiological update reports had been sent after August 16th, 2020, and all of these reports were made available on their official online platform.

The number of Covid-19 cases was still increasing at its highest rate since the beginning of the pandemic, with more than 5.7 million new cases reported per week, according to the most recent WHO weekly epidemiological update report, which was released on May 2, 2021, 10 am CET (Organization, 2021). This came after weekly reports of reported cases increased for nine weeks before. Over 93,000 more fatalities had been freshly recorded than the week before, making it the eighth week in a row. It was estimated that 3,186,817 of the 151,812, 556 reported cases of Covid-19 have led to fatalities. As of the date shown above, there have been 151,812, 556 documented cases of Covid-19.

According to the WHO's Covid-19 Dynamic Infographic Dashboard Iraq 2020-2021 scenario as of May 16th, 2021, there were 1,139,373 confirmed cases of the illness in Iraq as of that date. Sadly, 15,954 of these individuals have died from the condition, but luckily, 1,045,240 of them had also been cured. Iraq is a nation that is generally afflicted by the sickness on a global scale. This means that there were only 78,179 active instances, which was the conclusion that might be reached. It is crucial to note that the PCR test was administered to 9,872,873 distinct individuals. According to the WHO's Covid-19 Dynamic Infographic Dashboard Iraq 2020–2021 situation for May 16, 2021, there were 50,925 confirmed cases, of which 44,811 cases had been cured, and 1,142 dead cases had been recorded, leaving only 4,972 active Covid-19 sufferers (Organization, 2020b). This is due to the fact that this work is intended to be implemented on a sample drawn in this city.

The KRI authority quickly put in place preventive measures to stop the virus from spreading in the KRI as a whole. By the middle of March, all of the borders between the Kurdish region and the rest of Iraq and its neighbours (Iran on February 21, Syria on March 1, and Turkey on February 29, 2020) had been closed. On March 13, a total closure went into effect, making it hard to move between and within governorates and making it illegal to travel by land or (Thye et al., 2022; Ward et al., 2021). It's important to remember that these measures worked very well because, as of the end of May 2020, there had only been 606 confirmed cases and six deaths in the area.

Iraq had a high score on indicators of global, political, social, environmental, and security fragility. OECD (2020) reported on fragility show that Iraq was not ready for a number of big problems before the Covid-19 crisis (Bank, 2020). In fact, many experts pointed out before the epidemic that Iraq had all the signs of a weak dictatorship that was falling apart, not just a weak dictatorship. In 2019, Tony Cordesman "Iraq is on the verge of disaster," said a report from the World Bank that came out at the end of September. Even though the Iraq War started almost 20 years ago, the country is still insecure because of growing political unrest and division, diplomatic threats, growing civil unrest, and a growing gap between the government and the people (Verity et al., 2020; K. Wang et al., 2020). The effects of Covid-19 on public health and the economy had been especially hard on the displaced, women, and girls in Iraq. The government of KRI had taken a number of preventative and corrective steps to deal with the epidemic and its effects on other people. However, since early June, the number of confirmed Covid-19 cases has seen a staggering increase, leading to a sharp rise in the country's poverty rate. The situation has been exacerbated by a surge in job losses and increased spending, pushing the poverty rate to its peak.

Covid-19 infected people might not have any symptoms, had mild to moderate upper respiratory symptoms, or have severe pneumonia (N. Chen et al., 2020; Notari & Torrieri, 2022; Pijls et al., 2021). Consequently, individuals with severe infections face the risk of developing acute hypoxemic respiratory failure, a condition where their blood oxygen levels drop dangerously low, necessitating respiratory assistance. Such critical cases often require mechanical ventilation, but even with this intervention, there remains a high probability of severe illness or mortality. Depending on resource availability and the distribution of medical services, the death rate has been estimated to range from 1% to 10% based on patients' overall condition (Ranney et al., 2020).

This global health crisis has affected almost every nation, including Iraqi Kurdistan, where Covid-19 cases continue to rise. The impact of the disease is also evident in Europe and neighboring Iran, adding to the challenges faced by countries in close proximity.

Recently, there has been a significant surge in interest surrounding the application of non-parametric techniques to address real-world problems. Numerous studies have been conducted on neural network processes (NN) due to their enhanced flexibility compared to parametric models. These methods incorporate Bayesian learning, which can result in complex posteriors that are challenging to interpret.

The application of a multilayer perceptron neural network (NN) model proves beneficial in accurately estimating the occurrence rate of heart disease. This model operates on the principles of a mathematical representation of neuron behavior in the human brain. By simulating the brain's decision-making process based on historical data, the NN learns from a set of training examples, effectively creating a miniature replica of the brain's decision-making mechanism.

One of the most compelling aspects of NN is its ability to learn from examples (Radev et al., 2021; Ranney et al., 2020; Stouten et al., 2022; Ward et al., 2021). As a result, when utilizing an NN, there is no need to explicitly define how outputs are derived from specific inputs, as the network autonomously discerns the connections between inputs and outputs through its learning mechanism. This adaptability significantly impacts how the network is structured and the significance attributed to each connection. With sufficient training, the NN becomes adept at determining the output for any given set of input values and parameters.

Background Of the Study

The onset of the Covid-19 outbreak, a viral disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was initially detected in Wuhan, China in December of 2019. This novel infectious disease rapidly emerged as a significant global health concern. As of January 20, 2020, there were 282 confirmed cases of Covid-19 in China, Thailand, Japan, and the Republic of Korea. The virus was initially found in February 2020 in southern Iraq. After nine weeks of increases, Covid-19 cases were still at their highest rate since the pandemic began, with over 5.7 million new cases recorded each week. For the seventh week running, the number of reported deaths had exceeded 93,000. As per J. Wu et al. (2020), there have been 1,139,373 confirmed cases, 15,954 fatalities, and thankfully 1,045,240 recoveries. Iraq receives adverse ratings across various measures of political, social, environmental, and security instability worldwide. By mid-March, the Kurdish borders with Iran, Syria, and Turkey were all closed, along with the borders with the rest of Iraq. Additionally, on March 13, all forms of transportation, including those between and within governorates, as well as land and sea routes, were completely halted.

Neural network (NN) procedures are non-parametric ways of addressing problems that have broad practical applications. NN is a mathematical model that simulates, computationally, the activity of human brain neurons. The NN is made up of a number of processing parts linked together by weighted connections. Referencing to Ashish et al. (2004), the training method changes the weights of the connections in a way that reduces error. This means that the NN may understand the existing relationship between input and output with the help of a learning algorithm, eliminating the requirement for explicit programming. The study's overarching goal is to demonstrate how machine learning techniques may be applied to medical datasets to enable more nuanced analysis and informed decision-making. Since these patients were hospitalised after contracting Covid-19, their profiles are our primary focus. Age, sex, marital status, monthly income, residency type, country, number of hospital stays, status (cured/died), lab tests (including blood test, chest test, urine test, and etc.), and medications are only few of the many characteristics that wind up in the dataset. To conduct this research, we will utilise both classical mathematical methods and data mining techniques. Descriptive statistics will be used to characterize the structure and distribution of the study's variables.

Ramasamy et al. (2020) highlighted that as the number of dimensions in the input data increases, a larger amount of training data is required to achieve accurate classification. The Multilayer Perceptron, a fully connected, multi-layered, feed-forward supervised learning network, employs symmetric hyperbolic tangent activation functions. Back-propagation was employed to train the network and minimize quadratic error during the learning process.

Purpose of Study

Ultimately, the study's purpose is to demonstrate how machine learning techniques may be used to medical datasets in order to better detect correlations and interactions between different levels of variables. In addition, classic neural network may lead to overfitting when large number of inputs are entered in the model. In this study, we developed Bayesian neural network by involving Gibbs sampler while choosing prior values as well as hyperparameters which provided more accurate and fast convergence in the later stages of the HMC algorithm. Moreover, the contribution of the significant variables has not been pointed out in the current existed literature, thus all effective factors are stated in this study. The accuracy and effectiveness of multi-level analyses of variables defined by one of the measures utilized will be evaluated finally in order to determine factors that contribute to Covid-19.

Hypotheses. To address the risk factors, the fowling hypotheses were driven:

H01: Patient’s biographic has no effect on patient mortality by Covid-19 infections

H02: Pathological Test lab has no effect on patient’s mortality by Covid-19.

H03: Bayesian ANN has similar accuracy compared to ANN

H04: Logistic Regression with MCMC estimation (Bayesian Logistics regression) has the same accuracy in prediction

H05: Bayesian Neural Network has not out performed compared to other machine learning methods.

Research Questions

1. How to figure out the health factors that lead to the release of Covid-19 with a cured/dead result?
2. What will be the level of risk that Covid-19 poses to men and women?
3. Is machine learning algorithms and traditional statistical analysis methods suitable to predict the risk and pattern of Covid-19 mortality?

Significance of Study

The insights obtained from this research will play a crucial role in shaping health policies and legislation concerning women's healthcare within the country. Given the ongoing prevalence of the disease worldwide and the lack of a definitive cure, federal health agencies and governments will utilize these outcomes to develop effective strategies to improve women's health in healthcare institutions. Constructing a robust model based on the gathered information holds the potential to reduce the risk of mortality. Clinicians would gain valuable insights into the numerous risk factors contributing to maternal mortality, empowering them to make more informed decisions in their medical practices. Emphasizing evidence-based decision-making is a pivotal aspect of the medical field.

Beyond clinicians, this research holds significance for non-clinicians, including mental health professionals and psychologists. A better understanding of how behavioral and socioeconomic factors influence maternal health will enable them to address and solve these challenges more effectively. Ultimately, the research contributes to a holistic approach to enhance maternal healthcare and underscores the importance of data-driven decisions in the medical realm.

In addition, this research will provide valuable insights for non-governmental organizations (NGOs) and other activist groups dedicated to reproductive health. By gaining a deeper understanding of the disease and its impact on patients and healthcare providers, these organizations can develop innovative approaches to enhance patient education and healthcare monitoring. Furthermore, academics and researchers will benefit from the study's findings, as it offers a novel perspective for future research planning. The data generated can serve as a solid foundation for further investigations and advancements in the field. Moreover, individuals afflicted with Covid-19 will directly benefit from the study's outcomes, as it heightens awareness regarding the significance of comprehensive medical and emotional care for those affected by the disease. The results will foster a greater understanding of the importance of holistic support and reinforce the urgency to aid those battling Covid-19 with both medical expertise and empathetic care.

Limitation

Several limitations should be considered in this study. Firstly, there was a considerable amount of missing data in the laboratory and radiological records, which impeded their inclusion in the analysis. Secondly, the identified predictive factors may have been confounded by unmeasured variables, such as occupation, length of hospital stay, and pregnancy status. It is possible that medical staff and pregnant women had different disease severity profiles. Furthermore, no information was available regarding to if the patients wore mask face as precautionary measurements, this might have had great impact on the results. Additionally, the absence of reported medications may have impacted the disease status of the infected cases and could potentially lead to different conclusions.

CHAPTER II

LITERATURE REVIEW

Introduction

This chapter presents literature review of Bayesian statistical inference on artificial neural network analysis as well as Logistic regression analysis with Markov Chain Monte Carlo Techniques for Assessing Characteristics and Risk Factors of Covid-19. Chandra and He (2021) studied a piece of academic writing known as a literature review has the purpose of demonstrating the author's familiarity with and command of the collected corpus of scholarly work that is pertinent to a certain subject area. A critical examination of the works that were researched is included in a literature review, which is why it is not referred to as a literature report but rather as a literature review. A literature report does not contain this component. Therefore, this literature review will highlight how machine learning approaches can be applied to medical datasets to better distinguish correlations and relationships between various levels of variables in order to make more educated decisions. Finally, as per the research conducted by Ghoshal and Tucker (2020) precision and efficacy of multi-level analyses of variables as described by one of the measures used would be assessed in order to identify factors that lead to Covid-19.

The Significance of the Occurrence and Incidence of Covid-19

The study conducted by Morin et al. (2021), stated that despite the fact that a number of researches have been conducted on the topic of Covid-19's long-term effects, each of these studies has severe limitations. For example, a French telephone research with 478 patients and a response rate of 57% found that four months after patients were hospitalised for Covid-19, more over half of the patients had at least one symptom of long-COVID. 13 percent of participants in an app-based cohort study of 4,182 instances of Covid-19 self-reported long-COVID characteristics, and there was some suggestion that a greater frequency in long-COVID features was seen in women and older people.

According to Halpin et al. (2021), a significant proportion of hospitalized Covid-19 patients (63%) reported experiencing weariness or muscular weakness, while 26% reported difficulties with sleep, 23% reported feelings of concern or depression, and 23% observed a decrease in the frequency of myalgia and headache. However, it is important to note that these studies lacked a control group, limiting their generalizability. Furthermore, their focus was primarily on hospitalized patients or individuals who voluntarily participated in telephone surveys or utilized specific applications.

On a different note, Ghoshal and Tucker (2020) conducted a study investigating the use of drop-weights based Bayesian Convolutional Neural Networks (BCNN) to estimate uncertainty in Deep Learning solutions. The aim was to enhance the diagnostic performance of human-machine teams using a publicly available Covid-19 chest X-ray dataset. The findings demonstrated a strong correlation between uncertainty in predictions and prediction accuracy. The authors believe that the availability of uncertainty-aware deep learning solutions could facilitate broader adoption of Artificial Intelligence (AI) in clinical settings.

Taquet, Dercon, et al. (2021) highlighted numerous studies which have drawn comparisons between the long-term effects of Covid-19 and influenza, as highlighted in the literature. In a retrospective cohort analysis spanning six months and involving 236,000 Covid-19 patients, based on electronic health records (EHRs), higher occurrences of anxiety, mood disorders, sleeplessness, and dementia were observed following Covid-19 compared to influenza. Another study utilizing EHR data from American veterans (88 percent of whom were male) revealed that Covid-19 was associated with a higher prevalence of complications across various body systems when compared to influenza.

However, there is currently a lack of systematic estimates regarding the incidence and co-occurrence of long-Covid features, their correlation with age, sex, illness severity, and the relative risk compared to influenza within a large population. Recognizing this knowledge gap, the authors sought to address these issues utilizing EHRs. They conducted an analysis of the co-occurrence network, employing score-matched propensity of patients, Kaplan-Meier analysis, and Cox proportional hazard models. These methods were employed to gain insights into the aforementioned aspects and provide estimates that are currently lacking in the literature.

The Level of Risk Factors of Covid-19

In a retrospective cohort analysis conducted by Taquet, Geddes, et al. (2021), connected electronic health records (EHRs) of 81 million individuals were utilized, including 273,618 Covid-19 survivors. The study aimed to determine the incidence and co-occurrence of nine primary long-Covid features. These features encompassed various symptoms such as trouble breathing or shortness of breath, exhaustion or malaise, chest or throat discomfort, headache, stomach symptoms, myalgia, other pain, cognitive impairments, anxiety or depression, and other forms of pain. The calculations were performed within a three to six-month period following the diagnosis of Covid-19. Furthermore, the study examined the network of co-occurrence among these features. A comparison was made with a group of patients whose propensity scores were matched based on their influenza diagnosis during the same time period. Kaplan-Meier analysis and Cox proportional hazard models were employed in this comparative analysis.

It is worth noting that the study by Jalali et al. (2020) supports the aforementioned comparison, utilizing similar methods such as Kaplan-Meier analysis and the Cox proportional hazard model. This investigation was conducted over a specific time period, with the prevalence of atopic dermatitis being explored as a control group for reference.

Taquet, Geddes, et al. (2021) highlighted that among Covid-19 survivors (mean [SD] age: 46.3 [19.8], 55.6% female), 57.00% had one or more long-COVID feature recorded during the whole 6-month period (i.e., including the acute phase), and 36.55% between 3 and 6 months. The acute phase of Covid-19 is characterised by a sudden onset of flu-like symptoms that can last for days or weeks. The survivors of Covid-19 had an average age of 46.3 years old with a standard deviation of 19.8 years. The percentage of people who experienced each symptom was as follows: abnormal breathing (18.71 percent in the 1- to 180-day period and 7.94 percent in the 90- to 180-day period), fatigue/malaise (12.82 percent and 5.87 percent, respectively), chest/throat pain (12.60 percent and 5.71 percent), headache (8.67 percent and 4.63 percent, respectively), other pain (11.60 percent and 7.19 percent), abdominal symptoms (15.58 percent and 8.29 percent), myalgia (3.24 percent (22.82 percent ; 15.49 percent). According to Taquet, Dercon, et al. (2021) following Covid-19, each of the nine features was reported more frequently than following influenza (with a total excess incidence of 16.60 percent and hazard ratios ranging from 1.44 to 2.04, all of which were significant at the 0.001 level), they co-occurred more frequently, and they constituted a more interconnected network.

According to Wang et al. (2021), notable variations in incidence and co-occurrence were observed based on factors such as age, gender, and disease severity. In addition to the inherent limitations associated with the use of electronic health record (EHR) data, there are several additional limitations to consider in this study:

1- Generalizability: The findings may not apply to individuals who had Covid-19 but were not formally diagnosed, or to those who did not seek or receive medical attention for long-Covid symptoms. The study's scope is limited to patients who were identified through medical channels.

2- Persistence of Clinical Features: The study does not provide information regarding the duration or persistence of the observed clinical features. It focuses on the incidence and co-occurrence within a specific timeframe, without tracking the long-term trajectory of symptoms.

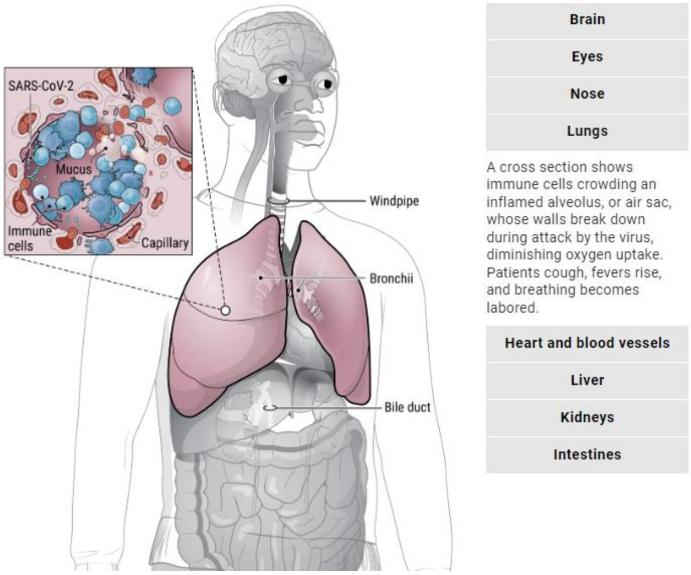
3- Cohort Discrepancies: Differences observed between cohorts could be influenced by variations in medical attention-seeking behaviour. One cohort may have sought or received more medical attention, potentially impacting the observed outcomes.

Covid-19's Impact on Bodily Organs

When individuals contract SARS-CoV-2 and become clinically ill, the respiratory system is often the primary organ affected. However, the virus has the potential to impact any organ in the body. Critical illness can result in damage to multiple organs. The virus attaches to angiotensin converting enzyme 2 (ACE2) receptors that are present in various tissues including vascular endothelial cells, lungs, heart, brain, kidneys, intestines, liver, pharynx, and others (Jain, 2020). This attachment can lead to direct injury to these organs, while systemic complications caused by the virus can also contribute to organ dysfunction. It's important to assess for damage to multiple organs when treating patients, as disturbances in coagulation and vascular endothelium can cause injury to multiple organs, even if they don't result in symptoms during the early stages. Among non-surviving patients, cardiac and renal dysfunction are frequently observed. Additionally, it is worth noting that organ injuries may not manifest immediately and could become apparent after the acute infection has resolved. Different organs may be affected at distinct time points, and there is a possibility of long-term chronic injury. As a result, the process of rehabilitation can be prolonged and demanding.

Covid-19 can trigger an excessive release of cytokines, leading to systemic inflammation, multi-organ injury, and potentially fatal outcomes (Merad & Martin, 2020). This immune response is referred to as cytokine storm or hypercytokinemia. Endothelial cells in various organs can be infected by SARS-CoV-2, resulting in vasoconstriction, inflammation, hypercoagulability, and edema, which in turn can cause organ ischemia (Carsana et al., 2020). According to F. Wang et al. (2020), the inflammatory response may persist even as the viral load decreases. Importantly, individuals with pre-existing immune-mediated inflammatory diseases who receive ant-cytokine biologics and other immunomodulatory treatments are not at an elevated risk for Covid-19. For a visual representation, please refer to Figure 1, which illustrates the affected parts of the human body in relation to Covid-19.

Figure 1:
Human's Body with description of Covid19 Impact



A cross section shows immune cells crowding an inflamed alveolus, or air sac, whose walls break down during attack by the virus, diminishing oxygen uptake. Patients cough, fevers rise, and breathing becomes labored.

Source:

https://www.science.org/content/article/how-does-coronavirus-kill-clinicians-trace-ferocious-rampage-through-body-brain-toes

Effect on coagulation

Covid-19 is associated with a range of blood clotting disorders, including deep vein thrombosis, pulmonary embolism, and disseminated intravascular coagulation (Bikdeli et al., 2020; Creel-Bulos et al., 2020). The risk of clotting is amplified by factors such as inflammation, hypercoagulability, endothelial dysfunction, blood vessel constriction, hypoxia, and immobility. To mitigate these risks, guidelines recommend thromboprophylaxis using medications like low-molecular-weight or regular heparin, fondaparinux, or direct oral anticoagulants such as apixaban or rivaroxaban. It's worth noting that heparins can also regulate the cytokine interleukin-6 and reduce immune activation. Following hospital discharge, extended prophylaxis may be beneficial (Wichmann & Sperhake, 2020).

According to Zhang et al. (2020), fever and inflammation in Covid-19 contribute to hypercoagulability and impaired fibrinolysis. Elevated levels of interleukin-6 (IL-6) are associated with both hypercoagulability and disease severity. Thrombosis is linked to increased antiphospholipid antibodies. Furthermore, Covid-19 stimulates the liver to produce procoagulant substances and moderately prolongs prothrombin time and activated partial thromboplastin time. It also leads to moderate thrombocytopenia, elevated C-reactive protein, lymphocytopenia, increased D-dimer levels, elevated fibrin degradation products (FDPs), and disseminated intravascular coagulation (DIC). D-dimer levels and DIC serve as prognostic indicators in Covid-19.

Effects on Pulmonary

Autopsy studies have revealed that patients in the acute phase of Covid-19 display characteristic diffuse alveolar damage, characterized by the absence of organization and fibrosis. This damage is attributed to the disruption of endothelial and alveolar cells, leading to the leakage of fluid and cells, as well as the formation of hyaline membranes (Barton et al., 2020).

Cardiac Effects

Cardiac complications associated with Covid-19 can manifest independently of pulmonary and other issues (Akhmerov, 2020; Madjid et al., 2020). Individuals with preexisting coronary artery disease (CAD), latent CAD, or no CAD are at risk of experiencing ischemic cardiac injury due to plaque rupture, thrombosis, or insufficient oxygen supply. Antiplatelet and anticoagulation therapies may be beneficial for managing acute coronary syndrome resulting from plaque rupture, while fibrinolytic therapy and percutaneous coronary intervention may be considered. It has been reported that the incidence of acute myocardial infarction has decreased during the Covid-19 period. In some patients, the virus can invade myocytes, while systemic inflammation such as cytokine storm can lead to myocarditis. These conditions can contribute to heart failure and arrhythmias, even after the acute phase of the infection has subsided and in the absence of lung damage.

Effect On Brain

Covid-19 has been found to depress brain stem reflexes, including the one responsible for detecting oxygen deprivation. Neurological symptoms can manifest as the primary symptoms or alongside respiratory or other manifestations, and they are more prevalent in severe cases of the disease. These symptoms, which include dizziness, headache, altered consciousness (such as confusion and delirium), and difficulty awakening, may be attributed to abnormal levels of oxygen and carbon dioxide. Delirium, a common neurological symptom, can lead to long-term cognitive impairment and memory deficits. The scarcity of commonly used sedatives has resulted in the use of benzodiazepines for sedation, which may worsen delirium. While hypoxic changes are observed in the brains of deceased patients, encephalitis or other viral-induced alterations are rare (Solomon et al., 2020).

Effect On Eyes

Referencing to Colavita et al. (2020), cells present on the ocular surface, including those in the cornea, inside the eyelids, and in the white of the eye, possess ACE2 receptors and TMPRSS2 proteases, which are necessary for the infection of SARS-CoV-2. Ocular abnormalities, such as conjunctivitis, are experienced by approximately one-third of hospitalized patients, with a higher incidence observed in those with more severe illness. Ocular involvement can manifest early in the disease progression, and cells on the ocular surface serve as potential entry points and reservoirs for the virus. It is important to note that the virus can be shed in ocular secretions, contributing to transmission, and it can remain infectious in the eye for up to three weeks.

Effect On Skin

Skin manifestations in Covid-19 share similarities with other viral infections and chronic inflammatory conditions such as eczema, acne, rosacea, and psoriasis. Vascular complications associated with these skin manifestations can have various causes, including neurogenic, microthrombotic, or immune complex-mediated mechanisms. The majority of patients with skin symptoms present with a patchy erythematous rash, while others may experience hives or widespread urticaria. In some cases, individuals may develop vesicles or blisters resembling those seen in chickenpox or rashes similar to measles. These skin lesions are primarily observed on the trunk, and itching is usually mild or absent. Skin eruptions can occur at the onset of symptoms or during hospitalization and typically resolve within a few days. It's important to note that the presence of skin manifestations does not indicate a more severe form of Covid-19.

Psychological Effects

The financial challenges and social isolation brought about by Covid-19 can contribute to a range of psychological issues that may persist even months after the initial outbreak. "Deaths of despair," including substance abuse and suicide, have seen an increase during the pandemic, particularly among individuals with dementia, mental illness, and autism. To address these concerns, it is important for individuals to engage in communication, either in-person or online, with friends and support professionals. Following discharge from the intensive care unit (ICU), approximately one-third of patients may experience dysexecutive syndrome, which can lead to symptoms such as inattention, disorientation, or difficulties with organized movements in response to commands. Furthermore, some individuals who have recovered from Covid-19 may develop mental health issues such as anxiety, depression, and post-traumatic stress disorder (PTSD). There is also a possibility of long-term effects, including an increased risk of developing Alzheimer's or Parkinson's disease (Dong et al., 2022).

Machine Learning Algorithms, and Traditional Statistical Analysis Methods

According to Jalali et al. (2020)’s study, due to the complexity of this treatment, there is a significant risk of death associated with it. The Norwood surgical technique offers the potential to restore functional systemic circulation in new born children who have congenital heart abnormalities involving a single ventricle. This study aims to fulfil the requirement for an accurate prediction of patient-specific risks related to one-year postoperative mortality, heart transplantation, and extended hospital stays. The study of Bucholz et al. (2018) aimed to assist medical professionals and the families of patients in making preoperative decisions. Either patient-specific risk variables are not taken into account in the risk prediction algorithms that are now available or the algorithms' only purpose is to estimate in-hospital death rates. The best solution is not either of these two options. They utilised data from the Pediatric Heart Network Single Ventricle Reconstruction project in conjunction with machine learning algorithms in order to assess and analyse each individual patient's risk of death as well as the length of time they will need to be hospitalised. After using a Markov Chain Monte Carlo simulation to complete some of the data that was missing, we included the results of that simulation together with the chosen variables into several machine learning models. Following that step, conclusions were drawn using the models. The deep neural network model used in this study demonstrated exceptional accuracy in predicting an individual's likelihood of death or requiring a heart transplant, achieving an accuracy rate of 89.4 percent and an impressive area under the receiver operating characteristic curve (AUROC) of 0.95±0.02. Additionally, for predicting longer hospital stays, the model showed a high accuracy rate of 85.3 percent and an AUROC of 0.94±0.04. These reliable prediction models and calculators can greatly assist in informed decision-making in both clinical and organizational settings.

As per the study by Dogan et al. (2021), infections with Covid-19 have prompted efforts all across the world to control and manage the virus, and maybe stop its transmission altogether. ML is a powerful tool that may be used in research on the Covid-19 virus as well as in the fight against it. This is a subject that is currently being researched. It is vital to keep up with the number of surveys that are emerging in the literature in order to stay up with the number of papers that are being published on Covid-19-related ML applications. In this study, we discuss recent discoveries that are associated with Covid-19 ML approaches.

We are focusing on the ability of machine learning to utilise clinical and laboratory data that is available to the public in order to make a diagnosis of Covid-19 and to make predictions about the risk of death and the severity of the disease. Sarker (2021) stated that analysing training data sets and different types of algorithms to better understand algorithms and feature selection supervised learning makes up the vast bulk of the machine learning strategies that are used in these two application cases. Due to the fact that the bulk of the relevant research consists of experiments, there has not been any application of previously developed models to the actual world as of yet. The ML models have diagnostic and prognostic properties that are consistent with the existing medical literature.

According to Ghoshal and Tucker (2020), machine learning (ML) is an area of artificial intelligence that aims to build self-sufficient systems capable of gaining knowledge from previous experiences. Ever since it was first introduced, the field of study known as machine learning has garnered a lot of interest due to its potential to solve a broad variety of issues that are encountered in the real world. Unsupervised learning, supervised learning, and reinforcement learning are the three main types of machine learning methodologies. In the process of supervised learning, an algorithm is given the opportunity to learn from a data set that has already been labelled. As per the study conducted by Fong et al. (2020), classification and regression are the two supervised learning approaches that are utilised the most often. On the other hand, unsupervised algorithms make an effort to learn from data that is not labelled. Data that has not been categorised is sent to the algorithms, and from this data, traits and patterns are extracted. Large-scale data sets with a high number of dimensions may be used for unsupervised machine learning techniques such as clustering and dimensionality reduction. According to the study by Cortés-Martínez et al. (2022) the algorithms used in reinforcement learning are designed to make mistakes and then to learn from those failures. As a result of this, a system of incentives and punishments is used while the individual is being trained.

MCMC in Neural Network Modelling

According to Jalali et al. (2020), MCMC techniques are commonly employed to tackle integration and optimization problems in high-dimensional spaces. These problems are crucial in various fields such as machine learning, physics, statistics, econometrics, and decision analysis. A recent survey has recognized the Metropolis algorithm as one of the top ten algorithms that have significantly influenced the advancement and practical application of science and engineering in the 20th century. This algorithm belongs to a broad category of sampling algorithms known as Markov chain Monte Carlo (MCMC). Over the past two decades, these algorithms have played a substantial role in statistics, econometrics, physics, and computer science (Chai et al., 2022). The authors emphasize that MCMC simulation remains the only known general approach capable of providing a solution within a reasonable time for various high-dimensional problems, including the computation of the volume of a convex body in multiple dimensions. The proof of a model $p(H|y)$ for ANNs (and other complex models) is theoretically intractable, necessitating multiple methodologies to calculate these possibilities for competing models. To approximate the data, one method uses Markov chain Monte Carlo (MCMC) calculations from the posterior weight distribution $p\ (w|y, H)$.

Monte Carlo Simulation for Epidemics

As a direct result of an improved comprehension of the detriment epidemics pose to public health and the economy on a global scale, there has been an appreciable rise in the frequency of MC in epidemic modelling estimates. This rise in MC frequency has been accompanied by an increase in the overall number of MC estimates. It gives decision-makers access to more complex probability statistics in the form of risk factors, which they may use to evaluate the options and the risk that is associated with them. For many decades, one of the most pressing problems in this area has been the statistical modelling of epidemic behaviours using MC. One of the pioneering groups in this area is Biazzo et al. (2022) who have been working on the development of a mathematical theory of epidemics from the year 1957. Using MC simulation techniques, Andersson and Britton analysed the behaviour of stochastic epidemic models and found the statistical features of these models for the next century. The purpose of this research was to gather information that may be used to stop future outbreaks.

Niraula et al. (2022), used a model that was quite close to the MC model in order to conduct their analysis of the level of population instability. Because of this, the researchers were forced to draw the conclusion that variations in the number of infected patients led to differences in the need of receiving emergency care. This is because of the idea that epidemiological indices, such the number of calls to emergency services, hospital admissions, and utilizations of intensive care units (ICUs), are vulnerable to change. A stochastic model of EMS occurrences and changes in demographic data is created with the use of empirical data from the Emergency Medical Services (EMS) facility in Poland's Lower Silesia Region. Because of the unpredictability of changes (in population numbers as people move out and an increase in the number of infected cases), the less-structured model cannot be examined using deterministic analytic methods.

Biazzo et al. (2022) states that the objective of this data synthesis is to ensure that only the most relevant information pertaining to the start and end of the Covid-19 epidemic is included. Within the context of this demonstration, MC will replicate the daily budget that must be allocated in order to prevent the further spread of a disease. This is what is in store for the future. The decision to use a composite model was made after considering the following factors: 1) the manner in which a person becomes infected is dependent on the intensity of travel (within a community, between cities, or internationally); 2) the prevalence of preventative measures; 3) the trail tracking of the suspected and quashed cases; and 4) the intensity of travel. The decision to use a composite model was made after considering the following factors: 1) the manner in which a person becomes infected is dependent on the intensity of travel (within a community). In theory, if the MC has access to more relevant data, it will perform better and offer more accurate solutions. This is because it will be able to better understand the context of the problem.

Neural Networks for Forecasting

As per the study by Khan et al. (2022), Deep learning neural networks need a number of processing layers in order to represent data at a high degree of abstraction. The term "deep learning" refers to the framework that is used by these networks in order to learn. Deep belief networks, long short-term memory networks, recurrent neural networks, and convolutional neural networks are only a few examples of the types of complicated machine learning models that fall under the umbrella term "deep learning" (LSTMs). In recent years, the study by Biazzo et al. (2022) has approved that the advances in computer power and the availability of enormous datasets, deep learning models have demonstrated outstanding performance in a variety of domains, including sentiment analysis, image analysis, and natural language processing. These achievements have been made possible by the advent of deep learning.

According to the study by Niraula et al. (2022), the Bayesian neural networks have a lot of promise for use in forecasting because of the promising accuracy of prediction that comes along with uncertainty quantification. Bayesian neural networks provide a lot of potential for application in forecasting. Both recursive Bayesian recurrent neural networks and evolutionary Markov chain Monte Carlo (MCMC) are types of Bayesian neural networks. Both of these types of networks are recurrent neural networks.

Covid-19 Model Complexities

According to the study by Niraula et al. (2022), the Covid-19 pandemic may be understood, at its most fundamental level, as an open complex system that has a significant amount of system complexity. They interact and relate to their environments and contexts in complex ways; they infect individuals and communities in unique ways; and they have significant emergent consequences and impacts on society in nearly every region of the globe. The above analysis has been approved in the study conducted by Khan et al. (2022) in which it has been stated that the virus, the disease, and their respective developments and transmissions each have a hidden nature as well as a high degree of uncertainty, self-organization, dynamics, and evolution. Despite this, the constrained and sparse data that are publicly available from Covid-19 do not clearly reveal the complexity and underlying epidemiological features, transmission mechanism, and cause-and-effect linkages indicated above. It is challenging to construct models that are accurate, durable, benchmarkable, and generally favourable when dealing with a limited amount of data. Levashenko et al. (2021), states that reaching aggressive modelling objectives with a small amount of data that is of poor quality from Covid-19.

CHAPTER III

Methodology and Statistical Learning of Bayesian Inference

Substantial theoretical background will be studied along with intensive methodology of the machine learning approaches in this chapter. Primarily, a Bayesian approach to solve regression problems will be presented, followed by an argument on parameter estimation in numerous nonlinear models. Furthermore, classical methods to determine joint confidence regions will be explored. Also, the chapter will cover a examination of Markov Chain Monte Carlo (MCMC) techniques and their implementation considerations., It describes the relevant concepts of artificial neural networks in a frequentist setting network as well as Bayesian neural networks and their way of implementing. We discuss appropriate regularization techniques, network architectures, and activation function ns. Finally, brief introduction of the applied models will be introduced.

Methodology

This chapter presents the patient's profile and methodology, including medical history, laboratory results, the person's demographics and etc. No personal information about the patients is explored such as, names, phone numbers, or addresses. Therefore, there is no need for a consent form or ethical paper to be signed by the patients because the dataset is completely anonymous. The process of discovering patterns, correlations, changes, deviations, and statistically significant structures and events in large datasets is considered. Distributions that characterize an observable property (descriptive statistics) are generated by classical statistical methods and used to assess the validity of a sample drawn from a larger population. To optimize Neural Network calculations, we present a novel metric that incorporates MCMC approaches and then evaluates the results against the standard approach.

Research Method

Site of study

We have used secondary data of a cross-sectional study type since the information was already collected by the hospitals itself. The data includes patients from 2020 to 2021.

SPSS Version 25 software, Python as well R code-based programming language and writing materials.

Study design

This is a quantitative cross-sectional descriptive analysis using a summary of patient files in the presentation. The quantitative approach is chosen because it is possible to interpret the information obtained for the statistical significance of associations between risk factors of Covid-19 and independent variables.

Study population

The target population comprised patients of Covid-19 take from the hospital in the recorded of 2020 to 2021. The study population comprised the Bayesian ANN analysis with MCMC approaches of Assessing Characteristics and Risk Factors of Covid-19 cases.

Data Capture and Analysis Strategies

Data Cleaning

In our data, there are so many missing participants; this missing information is the main key factor that can be used in the analysis. Also, we exclude the most extreme and unpredictable values from the data. We have also checked the consistency of the data. The data from patients’ files are captured in a Microsoft Excel spreadsheet which is $n = 537$.

Data Analysis

After cleaning and managing the data set, we have imported our data sheet on R software. All the data was carried out on R-Software for the analysis approach. We have calculated the frequencies and percentages of all variables. In bivariate analysis, we calculate the association of dependent variable Covid-19 with demographic, socioeconomic, and institutional variables using the chi-square and Cramer's v test. As we already defined that our dependent variable Covid-19 risk factors are in binary form, so that is the way we have to apply the logistic regression model, which helps us to describe the risk factors of patent’s characteristics influence the status of infected individuals in terms of duration of staying in hospitals, level of severity and our main point of discharged alive or dead among patients from two hospitals in private as well as public. Due to the association effects of these different risk factors, statistical approaches that analyse their multilevel and multi-layer relationships are needed to examine their effects. MCMC approach is appropriate to estimate the parameters. Our new proposed term is to involve MCMC techniques for optimization in Neural Network calculation and then compare it to default methods.

Variables Description

Table 1 shows the variable name, nature and coding briefly of the categorical variables only.

Table 1: Variable Descriptions of the Dataset

Variable Descriptions of the Dataset

Variable	Category	Code
Status	Died	1
	Recovered	0
Gender	Male	0
	Female	1
Age	Less than 18 years	1
	18 – 35 years	2
	35 – 65 years	3
	Greater than 65	4
Smoking	No	0
	Yes	1
Fever	No	0
	Yes	1
Cough	No	0
	Yes	1
Sputum	No	0
	Yes	1
Hypertension	No	0
	Yes	1
		1

Table 1 (Continued.)

Diabetes	No	0
	Yes	1
Stroke	No	0
	Yes	1

Statistical Methods and Techniques

In our study, we have applied descriptive statistics in which we calculate Mean, Median, SD, Quartiles. We have applied parametric and non-parametric comparison tests based on meeting normality assumption, Pearson Chi-Square, Cramer's V Sig, Odds Ratio, 95% CI (Low/high) for independence as well as comparison, and Logistic Regression (MLE) and Bayesian Logistic Regression. Our new proposed term is to involve MCMC techniques for optimization in Neural Network calculation and then compare it to default methods. Additionally, other machine learning approaches such as Random Forest, KNN, Naïve Bayesian and Support Vector Machine.

Validity

Validity refers to the capacity of a study to generate precise and meaningful outcomes that effectively capture the intended measurements. In this particular investigation, rigorous scientific research methods were employed to develop the data collection instrument and select the samples. The use of meticulously evaluated tools and techniques ensured that information bias was minimized, and the study's validity was upheld.

Generalizability

The results of the study are generalized to all the risk factors regarding the covid-19 cases. Beyond this inference could not be assumed.

Odds Ratio:

Odds ratio is the most commonly used in case-control studies; however, we can also be used in the cross-sectional study (Hasan et al, 2016). An odds ratio (OR) check the relationship between a disclosure and an outcome. The odds ratio instead of the odds that results will occur given a related revelation, compared to the odds of the outcome occurring in the absence of that exposure. For a probability of success π, the odds are defined to be:

$$odd = p / 1 - p$$

Confusion matrix.

The evaluation of a classification model's performance involves the use of a confusion matrix, which is a tabular representation. It effectively depicts the model's predictions by comparing them to the actual labels (James et al., 2013). This matrix, often organized as a 2x2 table (as shown in Table 2), presents the true class labels in the rows and the predicted class labels in the columns. As outlined by Raschka and Mirjalili (2019), the confusion matrix allows for the identification of four possible outcomes:

Table 2:

Confusion Matrix Exploration

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

- True positives (TP): The count of correctly predicted positive instances
- False positives (FP): The count that are incorrectly predicted as positive. (Also known as the Type 1 error).
- True negatives (TN): The number of instances that are correctly predicted as negative.
- False negatives (FN): The number of instances that are incorrectly predicted as negative. (Also known as the Type 2 error).

The confusion matrix can be used to calculate various evaluation metrics for a classification model, such as accuracy, precision, recall, and F1 score.

Accuracy

The accuracy metric indicates the proportion of correct predictions made by a model on the entire test dataset. It is commonly used as a primary evaluation metric for model performance. However, when dealing with imbalanced datasets, accuracy may not be a reliable metric to assess model performance.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.1}$$

Precision

Precision is a metric that quantifies the ratio of correctly predicted positive cases to the total predicted positive cases. It offers an assessment of the model's reliability. Precision is especially valuable when the cost of false positives outweighs that of false negatives.

$$Precision = \frac{TP}{TP+TN} \tag{3.2}$$

Recall (Sensitivity)

Recall is a metric that signifies the ratio of correctly identified actual positive cases to the total number of positive cases. It is a valuable measure when the cost of false negatives surpasses that of false positives. In these scenarios, recall provides an understanding of the model's capability to capture all positive cases without overlooking any.

$$Recall = \frac{TP}{TP+FN} \tag{3.3}$$

When the recall is high, the model can correctly identify most of the positive cases (TP+FN), resulting in a higher number of false positives (FP) and a lower overall accuracy. Conversely, a low recall indicates a higher number of false negatives (FN), implying that the model has missed identifying several positive cases. In such cases, the identified positive cases are likely to be true positives, thereby increasing the certainty of the model's predictions.

F1-Score

When precision is increased in a model, it usually leads to a decrease in recall, and vice versa. To provide a comprehensive evaluation of both metrics, the F1-score is employed, calculated as the harmonic mean of precision and recall. The F1-score achieves its highest value when precision is equal to recall. Therefore, the F1-score serves as a valuable metric for assessing the overall performance of a model, especially when precision and recall hold equal importance.

$$F1 - Score = \frac{1}{\frac{1}{Recall} + \frac{1}{Precision}} \qquad 3.4$$

Although the F1-score is a useful metric for combining precision and recall, its interpretability is limited. It is unclear whether the classifier is maximizing precision or recall when the F1-score is high. Therefore, it is often used in conjunction with other evaluation metrics to obtain a more inclusive thoughtful of the model's performance.

Receiver Operating Characteristic Curve

The Receiver Operating Characteristic (ROC) curve is a graphical representation that showcases the performance of a binary classification model. It illustrates the true positive rate (TPR) plotted against the false positive rate (FPR) at various threshold values. TPR, also known as sensitivity or recall, represents the proportion of true positive predictions, while FPR is the ratio of false positive predictions to the total number of actual negative cases. The ROC curve aids in determining the optimal balance between TPR and FPR for a specific classification problem.

The area under the ROC curve (AUC) is a widely used metric for evaluating the overall performance of a binary classification model. A higher AUC indicates a better ability of the model to distinguish between positive and negative cases (Davis & Goadrich, 2006). To plot the ROC curve, the following formulas need to be calculated:

$$False\ positive\ fraction\ =\ FP/(FP+TN)\ This\ is\ shown\ by\ X\ axis.$$

$$True\ positive\ fraction\ =\ TP/(TP+FN))\ This\ is\ shown\ by\ Y\ axis.$$

The Bayesian Logistic Regression Model

Introduction

A widely used approach to constructing prediction models for binary outcomes, particularly in medical research where the focus is on determining whether a patient has a disease or not, is through the use of logistic regression analysis.

In recent years, the Bayesian inference framework has gained popularity as a more appealing method for estimating parameters in logistic regression. This approach offers easier interpretability of parameter estimation and yields more dependable results for smaller samples. However, obtaining posterior distributions for the logistic regression parameters requires approximation methods as we cannot specify marginal posterior densities analytically. Therefore, we may utilize either Markov chain Monte Carlo (MCMC) approaches or Laplace’s method introduced by Tierney and Kadane (1986) for Bayesian posterior inference. In this research paper, we use Hybrid Monte Carlo approach with selecting proposal distribution using Gibbs sampler technique and compare with MLE estimator for predicting in-hospital death in patients with Covid19 disease.

Bayesian logistic regression offers several advantages, one of which is the ability to incorporate prior information during the modeling process. This integration of prior knowledge can enhance the accuracy of predictions and provide more informative inferences. Another benefit is that Bayesian methods enable the quantification of uncertainty in coefficient estimates, which can aid in decision making and model selection.

However, it's important to note that Bayesian logistic regression can be computationally demanding and may require a higher level of expertise compared to classical logistic regression. Despite these challenges, the benefits gained from incorporating prior information and quantifying uncertainty make Bayesian logistic regression a valuable approach in many research and decision-making contexts.

To conduct the Bayesian analysis, it is essential to establish a joint prior distribution over the parameter space. In this study, we choose to utilize an independent normal prior distribution with a mean of zero and low precision for the parameters. This choice is made to address concerns related to subjective beliefs influencing predictions of in-hospital death. Previous studies by Wilhelmsen et al. (2009) and Ziemba (2005) have also employed normally distributed priors for the model parameters, providing further support for this approach. By adopting this independent normal prior distribution, we aim to ensure transparency and mitigate potential criticism regarding the specification of subjective beliefs in the prediction process of in-hospital death.

Bayesian Inference Techniques

Bayesian inference offers a valuable approach for merging expert knowledge, also known as prior beliefs, with data to generate posterior beliefs. Consequently, in the event of new data being gathered, the Bayesian framework can be utilized to revise current knowledge by incorporating the fresh data. This updating process can be repeated as additional data is accumulated in the future. The fundamental principle behind all Bayesian inference is Bayes' theorem, which has been extensively explored in literature by scholars such as (Bernardo & Smith, 2009; Greenberg, 2012; Ntzoufras, 2009; Press & Press, 1989).

In order to comprehend the utilization of the Bayesian approach in parameter estimation, let theta denote the vector containing k unknown parameters and let X represent the vector containing n observations.

$$\beta = (\beta_1, \beta_2, \ldots, \beta_k) \qquad 3.5$$

$$X = (x_1, x_2, \ldots, x_n) \qquad 3.6$$

Based on Bayes' theorem, the posterior probability distribution $P(\beta/X)$ can be expressed.

$$P(\beta/X) = \frac{p(\beta) * p(X/\beta)}{\int p(\beta) * p(X/\beta)\, dX} \qquad 3.7$$

The expression is comprised of the following components $p(\beta)$ represents the prior distribution of the parameter, $p(X/\beta)$ represents the likelihood of the data given the parameters, X denotes the normalization factor, and $P(\beta/X)$ represents the posterior distribution. As the denominator term in equation (3.7) remains constant, it can be dropped, resulting in a simplified equation:

$$P(\beta/X) \propto\ p(\beta) * p(X/\beta) \qquad 3.8$$

When dealing with regression problems, the data is known but the parameters are unknown. As a result, the probability of the data given the parameters, $p(X/\beta)$, can be expressed as a function of the parameters known as the likelihood function, denoted by $l(\beta/X)$. Consequently, equation 3.8 can be written in terms of the likelihood function.

$$P(\beta/X) \propto\ p(\beta) * l(\beta/X) \qquad 3.9$$

The aforementioned equation illustrates the attractiveness of Bayes' theorem from an statistical perspective, as it facilitates the integration of prior knowledge with the information garnered from acquired data. With successive experimentation, this method permits the continual revision of parameter information.

Despite the many benefits of the Bayesian approach for estimating parameters in non-linear models, its application to medical problems often necessitates numerical solutions. When dealing with complex likelihood functions in engineering applications, analytical expressions are not always feasible due to the complexity of the integration. In these situations, Markov Chain Monte Carlo (MCMC) methods are advantageous from a Bayesian perspective as they provide a numerical approach for integrating high dimensional, complex functions.

Effect of Prior Distributions

The prior distribution plays a fundamental role in Bayesian inference, as it shapes the posterior inference. When selecting a prior distribution, it is common practice to refer to the existing literature, where the use of a normal distribution prior is often favored. The specification of the prior mean and variance holds particular significance: the former represents a prior point estimate of the parameter of interest, while the latter quantifies the level of uncertainty surrounding this estimate. A low variance prior indicates a strong prior belief, whereas a high variance prior indicates greater uncertainty.

In situations where prior knowledge is lacking, non-informative or vague priors are employed to prevent undue influence on the posterior distribution. These priors are typically improper, meaning they have non-integrable, infinite integrals. However, improper priors can be employed as long as they yield proper posteriors, ensuring valid inference (Ntzoufras, 2009).

Logistic Regression Model: Review

Assuming that we have the Binary logistic regression model, which can be expressed as:

$$p = P(y = 1/X) = \frac{1}{1+e^{-X\beta}} \qquad 3.10$$

In the context of the statistical model being discussed, the response variable is represented by the vector y. The values of the predictor variables for subject i, denoted as X_i

where i takes values from $1\ to\ q$, are included in the model, along with an intercept term. The design matrix X is a matrix of size n by $(q+1)$, which includes the values of all predictor variables for each of the n observations. The regression coefficients, denoted by the vector β, have a length of $q+1$ and include the intercept term b_0 along with the coefficients for the q predictor variables, denoted as $b_1, b_2, \ldots, b_q$.

Errors need to be independent but not normally distributed. Particularly in epidemiologic research, logistic regression is an effective tool because it enables the simultaneous analysis of several explanatory variables while minimizing the impact of confounding variables. Nevertheless, researchers must focus on model construction, avoiding just feeding algorithms with raw data and moving on to outcomes according to (Bocco et al., 2006). Some challenging judgments regarding model development will entirely depend on the abilities and expertise of researchers in the subject.

Likelihood and log Likelihood Function

$$f(y/X,\beta) = \prod_{i=1}^{n} p^{y_i}\,(1-p)^{1-y_i} \tag{3.11}$$

$$logf(y/X,\beta) = \sum_{i=1}^{n} -y_i(p) + (1-y_i)log(1-p) \tag{3.12}$$

By replacing p with $\left[\frac{1}{1+e^{-X\beta}}\right]$ and doing some calculations, the above formal becomes

$$logf(y/X,\beta) = \beta X(y-1_n) - 1_n\left[\mathrm{lo\ g}\left(1+1+e^{-X\beta}\right)\right] \tag{3.13}$$

Prior Distribution Proposal Selections

The process of selecting a model for our data involves specifying a prior distribution for the unknown coefficient in the applied model. We begin by assigning a non-informative flat prior with a huge variance to all the parameters, with assuming a mean of zero. However, we also incorporate a prior distribution with mean zero and a small variance of 1 for all the unknown parameters, which influences the posterior distribution. In Bayesian analysis, precision is often used instead of variance, where a large variance is considered non-informative and a small variance is not perfectly flat. We choose a large variance of 10000 (10^4) for our non-informative prior. To assign a prior to each unknown parameter, we adopt a normal distribution with a mean of zero to define the prior distribution. Consequently, the prior distribution takes the shape of a normal distribution.

We propose a multivariate normal prior for β

$$\beta \sim N(0, \sigma_\beta^2)$$

Its probability density function log without constant terms is as followings:

$$log\ p\left(\beta/\sigma_\beta^2\right) = -\frac{1}{2}\ log\ \sigma_\beta^2 - \frac{\beta^T\beta}{2\sigma_\beta^2} \qquad 3.14$$

Deriving Posterior distribution

The posterior distribution for the coefficients β is computed by multiplying the likelihood function, as described in Equation (3.13), with the prior distribution mentioned in Equation (3.14). This yields the posterior distribution, which is expressed as:

$$f\left(\beta/X, y, \sigma_\beta^2\right) \propto\ f(y/X, \beta) * \ p\left(\beta/\sigma_\beta^2\right) \qquad 3.15$$

Thus, the log of the posterior distribution is driven as below

$$log\ f\left(\beta/X, y, \sigma_\beta^2\right) \propto\ \beta X(y - 1_n) - 1_n\left[\log\left(1 + e^{-X\beta}\right)\right] - \frac{\beta^T\beta}{2\sigma_\beta^2} \qquad 3.16$$

As a result, the gradient function of the leapfrog function can be written as

$$\Delta_\beta \log f\left(\beta, X, y, \sigma_\beta^2\right) \propto X\left(y - 1_n + \frac{e^{-X\beta}}{1+e^{-X\beta}}\right) - \frac{\beta}{\sigma_\beta^2} \qquad 3.17$$

The equation (3.14) indicates that the prior distribution utilized in the Bayesian logistic regression model does not fall into a conjugate family, ruling out the possibility of using a conjugate prior. Furthermore, the normalizing constant in the denominator of equation (3.7) cannot be computed explicitly, necessitating the use of simulation methods to derive the posterior distributions of the parameters. Markov Chain Monte Carlo (MCMC) methods are commonly employed to generate a Markov chain with a stationary distribution that aligns with the posterior distribution of the vector β.

Approximate methods in Bayesian inference

Bayesian logistic regression aims to estimate the posterior distribution of the model coefficients, considering both the observed data and prior beliefs regarding their distribution. Equation (3.17) clearly illustrates that it is a complex function of the parameters, necessitating the use of numerical methods to compute the marginal posterior, posterior moments, and predictive densities for each parameter of the model. Approximations can be obtained using techniques based on Markov Chain Monte Carlo (MCMC).

Markov Chain Monte Carlo: The Basics

MCMC (Markov Chain Monte Carlo) methods are a diverse set of computational techniques utilized to estimate integrals and produce posterior samples. In Bayesian analysis, MCMC algorithms are commonly employed to approximate the posterior distribution by generating simulated samples. The Metropolis-Hastings (MH) algorithm is a widely used principle in Bayesian analysis for generating posterior samples. A specific variant of the MH algorithm is the Gibbs sampler, which is often utilized in practice.

Metropolis-Hastings

The Metropolis-Hastings (MH) algorithm employs a proposal density $\boldsymbol{q}(\boldsymbol{\beta}^{\boldsymbol{Prop}}|\boldsymbol{\beta}^{\boldsymbol{t-1}})$ to determine the values of β^{t-1} in the chain. Here, β^{Prop} represents the proposed value for the next element in the chain, while the proposal density is conditioned on the preceding value β^{t-1}. Several proposal functions can be used, with random walk proposals being the most prevalent approach, but we propose Gibbs sampler as a new technique.

Thus, the basic Metropolis-Hastings algorithm is as follows: a candidate state x is generated at step t from the proposal distribution $\boldsymbol{q}(\boldsymbol{\beta}^{\boldsymbol{Prop}}|\boldsymbol{\beta}^{\boldsymbol{t-1}})$. The candidate state is then either accepted or rejected as the next state in the chain, with probabilities determined by the algorithm.

Algorithm (1): Metropolis Hastings

```
$Setting\ starting\ point\ for\ \boldsymbol{log\ f\left(\beta/X, y, \sigma_\beta^2\right)}$
$\boldsymbol{for\ t = 1\ to\ N}$
    $\beta^{Prop} = \boldsymbol{q}(\boldsymbol{\beta^{Prop}}|\boldsymbol{\beta^{t-1}})$
    $u = Rand.Uniform\ (0,1)$
    $\propto = min\left(1, \frac{\boldsymbol{log\ f\left(\frac{\beta}{X}, y, \sigma_\beta^2\right)^{Prop} q\left(\beta^{Prop}|\beta^{t-1}\right)}}{\boldsymbol{log\ f\left(\frac{\beta}{X}, y, \sigma_\beta^2\right)^{t-1} q\left(\beta^{Prop}|\beta^{t-1}\right)}}\right)$
    $Accept\ \boldsymbol{\beta^{Prop}}\ if \propto < u, \boldsymbol{\beta^t} = \boldsymbol{\beta^{t-1}}$
$End\ for$
$\boldsymbol{Return\ \beta^{(1)}, \beta^{(2)}, \dots \beta^{(N)}}$
```

Hamiltonian Monte Carlo Algorithm

By utilizing a guided proposal generation scheme, Hamiltonian Monte Carlo (HMC) enhances the efficiency of the MH algorithm. HMC achieves this by utilizing the gradient of the log posterior, which directs the Markov chain towards regions of higher posterior density where the majority of samples are taken. Consequently, a well-optimized HMC chain is capable of accepting proposals at a significantly higher rate compared to the traditional MH algorithm (Gelman et al., 1997). Further detailed explanations can be found in other sources, such as (Betancourt, 2017; Neal, 2012).

The Hamiltonian function, denoted as $H(\beta, p)$, is expressed as the sum of potential energy $U(\beta)$ and kinetic energy $K(p)$, where β and p are both in the real $k - dimensional$ space, i.e., $\beta, p \in \mathbb{R}^\wedge k$. Specifically, the expression is given as $H(\beta, p) = U(\beta) + K(p)$.

In MCMC applications in statistics, our primary objective is to generate θ from a specified distribution $f(\beta)$. To achieve this, we set the potential energy function as $U(\beta) = -log\ f(\beta)$. By doing so, the generated β values from the Hamiltonian function will adhere to the intended distribution. In terms of momentum, it is usually assumed that p follows a multivariate normal distribution with mean 0 and a user-defined covariance matrix $M, i.e., p \sim N_k(0, M)$. With this formulation, we possess/obtain.

$$H(\beta, p) = -\ log\ f(\beta) + \frac{1}{2} p^T M^{-1} p \qquad 3.18$$

As time progresses, HMC moves along paths that are controlled by the first-order differential equations, which are commonly referred to as the Hamiltonian equations.

$$\frac{dp}{dt} = -\frac{\partial H(\beta, p)}{\partial \beta} = -\frac{\partial U(\beta)}{\partial \beta} = \nabla_{\beta} log f(\beta)$$

$$\frac{d\beta}{dt} = \frac{\partial H}{\partial p} = \frac{\partial H(\beta,p)}{\partial p} = \frac{\partial K(p)}{\partial p} = M^{-1}p \qquad 3.19$$

Therefore, the resolution of the Hamiltonian equations becomes a critical stage in HMC simulation. Although Euler's method is a conventional approach for solving differential equations, it tends to accumulate errors, particularly after numerous steps, as noted by Neal (2012). In HMC, a larger number of steps may be required to guarantee that the new proposal is adequately distant from the previous sample's location. The leapfrog method, introduced by Ruth (1983), is a viable alternative to Euler's method for approximating the solutions to Hamiltonian equations. The leapfrog algorithm adjusts Euler's method by utilizing a discrete step size ϵ for p and theta independently, with a complete step ϵ in theta surrounded by two half-steps $\epsilon/2$ for p.

$$p(t + \epsilon/2) = p(t) + (\epsilon/2)\nabla_{\beta} \log f\big(\beta(t)\big)$$

$$\beta(t + \epsilon) = \beta(t) + \epsilon M^{-1} p(t + \epsilon/2) p(t + \epsilon) = p(t + \epsilon/2) + (\epsilon/2)\, \nabla_{\beta} \log f\big(\beta(t + \epsilon)\big) \qquad 3.20$$

Algorithm (2): Hamiltonian Monte Carlo

```
$\boldsymbol{Input\ (\beta^{(0)}, log\ f(\beta/X, y, \sigma_\beta^2), M, N, L, \epsilon)}$
$Setting\ starting\ point\ for\ log\ f(\beta/X, y, \sigma_\beta^2)$
$for\ t = 1\ to\ N$
     $p = Rand.noramal(0, M)$
    $\beta^{(t)} = \beta^{(t-1)}, \tilde{\beta} = \beta^{(t-1)}, \tilde{p} = p$
    $for\ i = 1\ to\ L$
        $\tilde{\beta}, \tilde{p} \rightarrow Leapfrog(\ \tilde{\beta}, \tilde{p}, \epsilon, M)$
    $End\ for$
    $\propto = min\left(1, \frac{\exp\ (\log f(\ \tilde{\beta}) - \frac{1}{2}\tilde{p}^T M^{-1}\ \tilde{p}}{\exp\ (\log f(\ \tilde{\beta}^{(t-1)}) - \frac{1}{2} p^T M^{-1} p}\right)$
    $with\ probability\ \ \propto, \beta^{(t)} = \tilde{\beta}\ and\ p^{(t)} = -\tilde{p}$
$End\ for$
$Return\ \beta^{(1)}, \beta^{(2)}, \dots \beta^{(N)}$
Run (LeapFrog Function)
```

Artificial Neural Network Analysis

Because of their advances in fields such as image recognition, natural language processing, and reinforcement learning, significant attention nowadays is on the usage of neural networks regardless of their advancements in optimization and learning algorithms. Nevertheless, there are several issues that have yet to be addressed such as overfitting which it easily tends to produce such phenomena regardless of how well they perform on some data sets, and eventually provide poor generalization. Finally, their model hyperparameters need a large number of the tuning.

A neural network is defined as a parametric approach which attempts to estimate the x-y mapping. $f: x \rightarrow y$, given a certain set of data $D = \{x_i, y_i\} \in (X, Y)$. The set of neural network parameters is identified as weights w, and the issue of determining the set of weights that best describes the mapping f is calculated by applying Maximum Likelihood Estimation (MLE) technique.

$$W_{MLE} = arg\ \underset{w}{max}(\log p(X/W) \qquad 3.21$$

Feed-forward Approach

Neural networks, known as adjustable nonlinear regression and discriminant models, as well as models for nonlinear dynamical systems and data reduction, consist of numerous interconnected "neurons" that perform computations in linear or nonlinear ways. These neurons are often organized into layers. When utilizing neural networks (NNs) for data analysis, it is crucial to distinguish between NN models and NN methods. Many NN models resemble or can be directly compared to popular statistical tests, such as generalized linear models, binomial regression, nonparametric regression, discriminant analysis, projection pursuit regression, principal components analysis, and cluster analysis, especially when the focus is on prediction rather than interpretation. These NN models have the potential to be highly practical. However, standard NN learning algorithms are inefficient as they are designed for parallel processing computers but are often implemented on simple serial machines like regular PCs. The neural network formula, known as feed-forward, involves a linear combination of independent variables, their weights, and the bias (or intercept) term for each neuron.

A feed-forward network is a linear perceptron generalization in which artificial neurons layers are packed together with non-linearities utilized between each layer. According to Hornik et al. (1989), the universal approximation theorem is the reason for the many successful applications of neural networks. It demonstrates that feed-forward neural networks with non-linear activation functions and at least one hidden layer can approximate universal functions.

$$Z = bias + W_1 * X_{1i} + W_2X_{2i} + \cdots W_mX_n \qquad 3.22$$

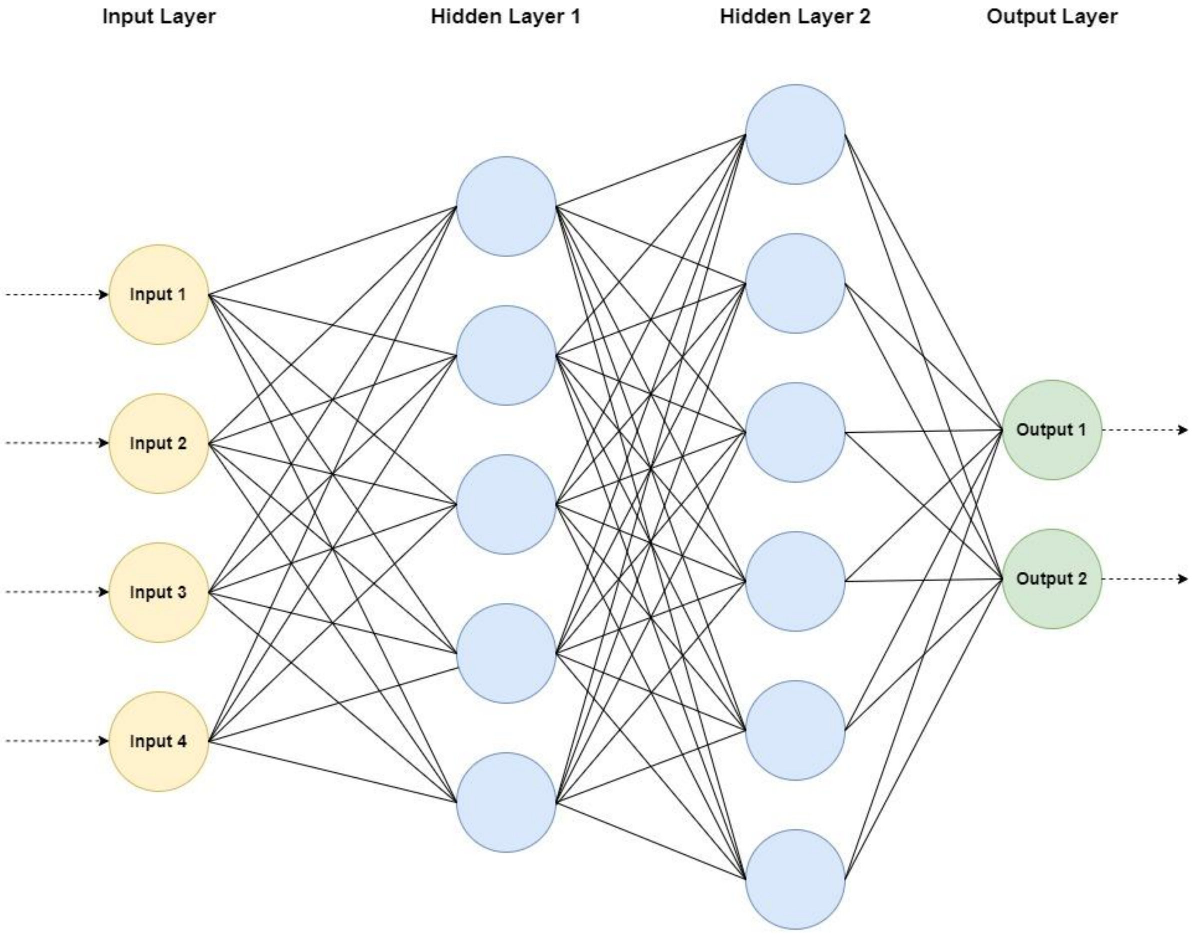
Where: Z is defined as the symbol for denotation of the graphical representation

Ws, are the weights or the beta coefficients

Xs are known as the inputs or independent variables

Figure 2:

Feed-forward neural network illustration with 2 hidden layers



By considering an input vector x and an output vector y, a feed-forward neural network can be constructed. This neural network consists of m hidden layers, each containing h_i nodes, along with weights w, bias b, and a non-linear activation function f. By utilizing this network, new outputs can be generated, as expressed by the following equation:

$$HL^{(1)} = \varphi(w^{(1)}x + bias^{(1)})$$

$$HL^{(2)} = \varphi(w^{(2)}HL^{(1)} + bias^{(2)})$$

$$\hat{y} = \sigma\left(w^{(3)}HL^{(2)} + bias^{(3)}\right)$$

In the case of classification, the function φ can be linear and σ can be non-linear activation functions, as is regularly the case for regression problems. There are numerous other options for φ *and* σ however.

Types Activation Functions

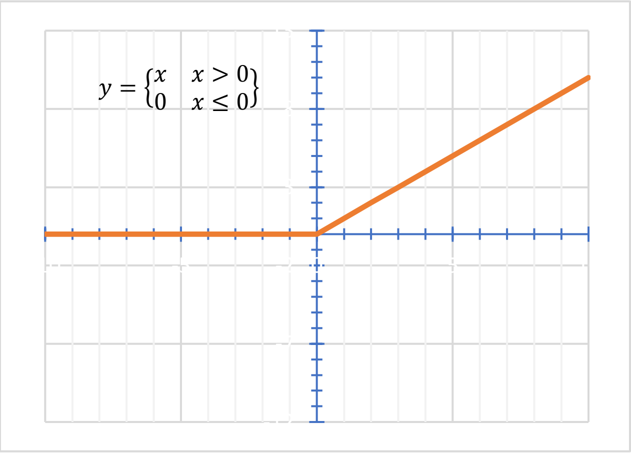
To empower neural networks in utilizing complex nonlinear transformations, some form of nonlinearity must be added to the model. There are numerous possible activation functions, but for the purposes of this thesis, we will focus only on two of them.

1- ReLU Activation:

This function is to be implemented for the input layer as well as hidden layer neurons.

Figure 3:

ReLU Activation Function Demonstration

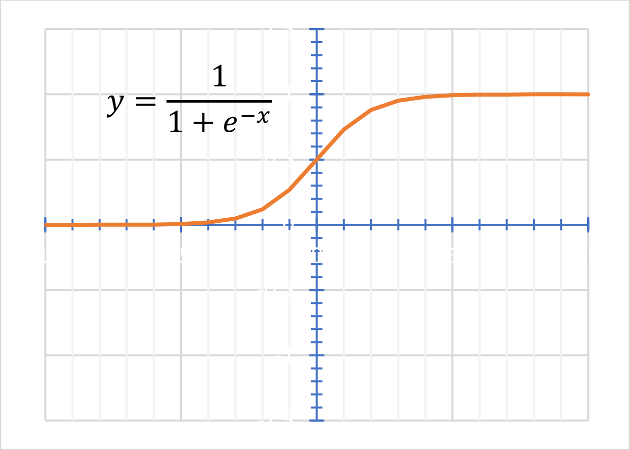


2- Sigmoid Activation Function

The sigmoid activation function is utilized with regressors because it "squishes" a set of outputs from 0 to 1 from negative infinity to positive infinity. The constraints denote the two classes that could exist. The sigmoid equation is as follows:

Figure 4:

SIGMOID Activation Function Demonstration



It will be run for the output values since we are predicting 0 and 1 values for the response variables.

Derivative of Sigmoid Function

Let us derive the Sigmoid function's derivative with respect to its input

$$\sigma = \frac{1}{1+e^{-x}} \rightarrow d\sigma = \frac{d}{dx}(1+e^{-x})^{-1} = -1 * \ (1+e^{-x})^{-1-1} * \frac{d}{dx}(1+e^{-x})$$

$$= -\ (1+e^{-x})^{-2} * (\frac{d}{dx}1 + \frac{d}{dx}e^{-x}) = -\ (1+e^{-x})^{-2} * (0 + e^{-x} * \frac{d}{dx}[-x])$$

$$= -\ (1+e^{-x})^{-2} * (e^{-x} * (-1 * \frac{d}{dx}x)) = -\ (1+e^{-x})^{-2} * (e^{-x} * (-1))$$

$$= -\ (1+e^{-x})^{-2} * (-e^{-x}) = (1+e^{-x})^{-2} * e^{-x}$$

$$= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x}) * (1+e^{-x})} = \frac{1}{1+e^{-x}} * \frac{e^{-x}}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} * \frac{1+e^{-x}-1}{(1+e^{-x})} = \frac{1}{1+e^{-x}} * (\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}})$$

$$\therefore d\sigma = \sigma * (1-\sigma)$$

To employ a neural network to learn from data, it first must be indicated what to learn. We are interested in determining the set of weights, w, that can fit to the training data while also generalizing to new data. As a result, we must develop a loss function that calculates the performance of how well our current model fits the data. Instead of calculating **categorical cross-entropy** for loss function, **binary cross-entropy** terminology is used and its partial derivative of the Binary Cross-Entropy loss is a fairly simple equation that will be straightforward to implement in practice.

$$Loss_i = (y_i)(-\log \hat{y}_i) + (1 - y_i)(-\log\,(\,1 - \hat{y}_i)$$

$$= -y_i * \log(\hat{y}_i) - (1 - y_i) * \log(1 - \hat{y}_i)$$

$$\frac{\partial Loss_i}{\partial \hat{y}_i} = \frac{\partial}{\partial \hat{y}_i}(-y_i * \log(\hat{y}_i) - (1 - y_i) * \log(1 - \hat{y}_i))$$

$$= -y_i * \frac{1}{\hat{y}_i} * 1 - (1 - y_i) * \frac{1}{1 - \hat{y}_i}(0 - 1)$$

$$= -\frac{y_i}{\hat{y}_i} + \frac{1 - y_i}{1 - \hat{y}_i} \rightarrow \frac{\partial Loss_i}{\partial \hat{y}_i} = -\left(\frac{y_i}{\hat{y}_i} - \frac{1 - y_i}{1 - \hat{y}_i}\right)$$

Regularization

Since neural network is highly complex and it normally ends with facing overfitting and thus produces very poor generalization while unseen datapoints go through to the model. There is a numerous regularization technique which have been proposed to overcome this issue.

1- Dropout

The addition of noise to the model is a different style to regularization. The idea is that by presenting noise during training, the model rather performs better on new data and will struggle to outperform fit. This concept has been applied in a variety of studies, such as adding Gaussian noise to weights or noise to training data. The most prominent approach, however, is dropout, as studied by Srivastava et al. (2014).

2- L1-Regularization

L1-regularization is often recognized as Lasso which penalize large weight's values where works on parameter Lambda as following:

$$Loss_i(y|x,w) + \lambda \sum_i^m \sum_j^{HL_i} |w_{ij}| \qquad 3.23$$

L is the loss function of the network for a new prediction, setting new datapoints, and trained weights. L1-regularization causes model sparsity by forcing parameters to equal zero. L1-regularization is equivalent to presenting a Laplace-prior on the distribution of weights in the network from a probabilistic standpoint.

3- L2-Regularization

L2-regularization, also recognized as Ridge regression, where large weights are forced to be penalized by adding an L2-penalty to the model's cost function as follows:

$$Loss_i(y|x,w) + \lambda \sum_i^m \sum_j^{HL_i} w_{i,j}^2 \qquad 3.24$$

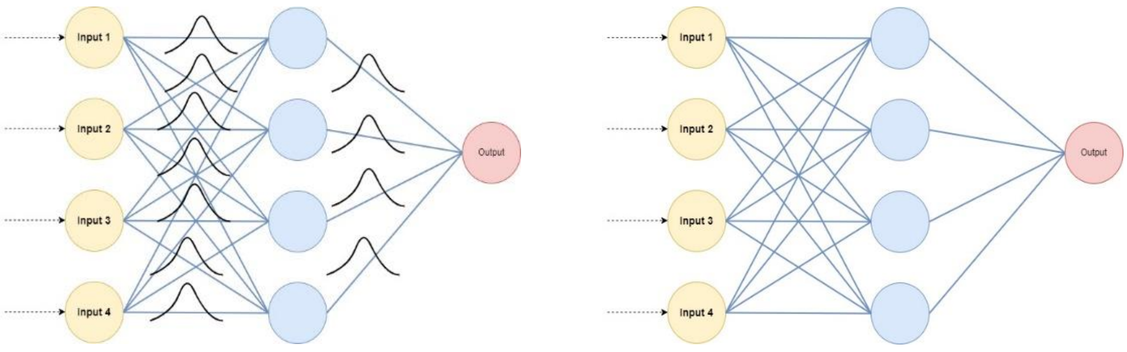
Bayesian Neural Network

Brief introduction

A Bayesian neural network (BNN) is a type of neural network that incorporates Bayesian inference into its architecture. Like traditional neural networks, a BNN consists of layers of interconnected nodes, or neurons, that receive inputs, apply weights and biases, and produce outputs. However, unlike traditional neural networks, a BNN is designed to model not only the mapping from inputs to outputs but also the uncertainty associated with that mapping. Figure 5 demonstrates the architecture of both Bayesian neural network as well as neural network with point estimation.

Figure 5:

Left: Bayesian Neural Network with probability distribution over weights. Right: Classic Neural Network with point estimates for weights.



In a BNN, the weights and biases of the neural network are treated as random variables with prior distributions. The prior distributions reflect the prior belief or uncertainty about the values of the weights and biases. The goal of Bayesian inference is to update the prior distributions based on observed data, resulting in posterior distributions that reflect the updated beliefs or uncertainty about the values of the weights and biases.

The use of Bayesian inference in a neural network has several advantages. For example, it allows for the quantification of uncertainty in the model's predictions, which can be useful in many applications, such as finance or medicine. Additionally, it can help prevent overfitting by regularizing the model and reducing the impact of outliers.

However, BNNs can be computationally expensive and require specialized algorithms to perform inference over the posterior distribution. Despite this, they have become an increasingly popular tool in machine learning and have shown promising results in various applications such as image classification, language modeling, and reinforcement learning.

Bayesian Learning Process

The Bayesian learning process begins with the definition of a model, M, and a prior distribution $p(w)$ for the model parameters a. After examining additional data used to update the updated distribution of priors is used to create the posterior distribution with utilizing Bayes' rule.

$$P(W/X) \propto p(W) * p(X/W) \qquad 3.25$$

Likelihood

In statistics, likelihood function is driven from its original pdf for n data points in multiplying the function for each joint (x, y) dataset. Its general written formula is as followings:

$$L(W/X) = \prod_{i=1}^{n} p(y^{(i)}|X^{(i)}, W) \qquad 3.26$$

In order to write in more detail explicitly for two class as in our cases where sigmoid function is intended to be used for the output layer, eq6 provides more insights on the likelihood function:

$$p(y = 1|X, W) = [1 + \exp(-f(x, w))]^{-1} \qquad 3.27$$

Free parameters where are called (weights) in the model must be set corresponding to the training set size, the noise level as well as the target function complexity when computing classical estimation (error minimization) for the MLP. However, limiting the network size is no longer an issue in the Bayesian technique, but it is wise to minimize number of hidden units in practice for computational purposes. In addition, referring Neal (2012), for small sample sizes, the process of converging tends to be Gaussian while implying limiting hidden unit numbers which is considered as a feasible practice in such circumstances.

Choosing Priors

Smoothing priors, for instance, state that functions with small values of second derivative are more likely in Bayesian approximation. According to (Bishop, 1993; Lampinen & Selonen, 1997), however, results in a fairly complex treatise with MLP. The MLP complexity can be roughly treated by varying the weights w size. Eventually, this can be accomplished, for example, by using a Gaussian prior distribution for weights w by specifying the hyperparameter alpha.

$$p(W|\alpha) = (2\pi)^{-m/2}\alpha^{m/2}\exp\left(-\alpha\sum_{i=1}^{m} w_i^2/2\right) \qquad 3.28$$

Vague hyperprior (α) is set due to the complexity as well as no knowledge for the correct value caused by the hyperparameter and this is attributable to produce either very high or very low values of alpha. Therefore, a Gamma distribution of vague type is expressed with mean and shape parameter alpha

$$p(\alpha) \sim GM(\mu, a) \propto \alpha^{a}\exp\left(-\alpha a/2\mu\right) \qquad 3.29$$

As Neal (1998) applied run separate priors each with its own α at each weight group from each layer in order to have a prior which is invariance under linear transformation of the dataset used. Neverthelss, this approach often cause creating common priors for all considered inputs and to avoid such phenomenon, Automatic Relevance Determination (ARD) is taken step into the discussion per (MacKay, 1994; Neal, 1998). The input-to-hidden weights linked to the identical input have shared variance extracted from the priors which lead to have same distribution of the prior (hyperprior). Consequently, prior’s posterior values are adjusted, and the weights of unrelated inputs move towards to zero.

The predictive distribution is achieved by calculating the model's integration of predictions regarding the posterior distribution to forecast the new output $y^{(n+1)}$ when new values of input $x^{(n+1)}$ is available.

$$p\left(y^{(n+1)}\middle|x^{(n+1)}, X\right) = \int p\left(y^{(n+1)}\middle|X^{(n+1)}, W\right) p(W/X)\, dW \qquad 3.30$$

This can be understood as taking the average of the predictions from all models, where the weights are determined by their respective posterior probability distributions. To illustrate, the predictive distribution for new data is derived by integrating the posterior distributions of the parameters and hyperparameters, resulting in the following expression:

$$p\left(y^{(n+1)}\middle|x^{(n+1)}, X\right) = \int p\left(y^{(n+1)}\middle|X^{(n+1)}, w, \alpha, \tau\right) p(w, \alpha, \tau/X)\, dw\alpha\tau \qquad 3.31$$

Various functions' expectations in relation to the posterior distribution for parameters is feasible to be evaluated and in regression, for instance, the probability of $y^{(n+1)}$ component can be evaluated as following:

$$\hat{y}^{(n+1)}{}_{k} = \int f_k\left(X^{(n+1)}, w\right) p(w, \alpha, \tau/X)\, dw\alpha\tau \qquad 3.32$$

Except in certain circumstances where Bayesian model is simple and both the likelihood and the prior are conjugate distributions, this integration is likely to be intractable. Thus, this is when numerical analysis such as Monte Carlo Markov Chain comes and its algorithms play major role in estimating model's posterior distribution for more complex models.

Hybird Monte Carlo Algorithm for BNN

The Hybrid Monte Carlo (HMC) algorithm is a computational method used in statistical physics, Bayesian statistics, and other fields to sample from complex probability distributions. The algorithm combines molecular dynamics simulations with Markov Chain Monte Carlo (MCMC) sampling to explore high-dimensional spaces efficiently.

In the HMC algorithm, the target distribution is represented as a probability density function (PDF), which can be evaluated up to a constant of proportionality. The algorithm starts by simulating a Hamiltonian system, consisting of a set of particles with mass and position variables, subject to a potential energy function. This simulation generates a new proposed state, which is then accepted or rejected based on the Metropolis-Hastings criterion.

The HMC algorithm is called "hybrid" because it combines the benefits of both Monte Carlo sampling and molecular dynamics simulations. The molecular dynamics simulation allows the algorithm to make proposals that are more efficient and less correlated than in traditional MCMC methods, while the Metropolis-Hastings criterion ensures that the algorithm produces valid samples from the target distribution.

Overall, the HMC algorithm is a powerful tool for sampling from complex distributions, particularly those with high-dimensional spaces or strong correlations between variables. It has been successfully used in a variety of applications, including Bayesian inference, machine learning, and computational physics.

Equation (3.32) represents the expectation of function $f(X^{(n+1)};W)$ with regards to the parameter's posterior distribution. The Monte Carlo method can implemtend to approximate this by drawing a sample of values $W^{(t)}$ from the posterior distribution.

$$\hat{y}_k^{(n+1)} \approx \frac{1}{N}\sum_{t=1}^{N} f_k\left(X^{(n+1)}, w^{(t)}\right) \qquad 3.33$$

The hybrid Monte Carlo (HMC) algorithm is employed to compute the parameters, while Gibbs sampling is utilized for the hyperparameters. HMC is an advanced Monte Carlo method that leverages gradient information to mitigate random walk behavior often observed in the Metropolis algorithm. The gradient provides guidance on the direction to explore in order to discover states with higher probabilities. The inclusion of Gibbs sampling for the hyperparameters helps to reduce the need for extensive tuning to achieve satisfactory performance in HMC.

Variational Inference

Variational inference is an alternative approach for approximating inference in Bayesian modeling that can be considered as a parametric alternative to the MCMC-sampling algorithm class. The advantage of variational inference is that it substitutes the integration component of inference with optimization or differentiation, which typically requires less computational resources. Nonetheless, the primary limitation of these algorithms is that they assume independence among the model's parameters and the resulting variational posterior may deviate substantially from the true posterior. In essence, variational inference presents a powerful option for replacing integration with optimization, but its assumptions and potential divergence from the true posterior must be taken into account.

The aim of variational inference is not to perform sampling from the exact posterior but rather to utilize a distribution $q_\theta(W)$, referred to as the variational distribution, that is parameterized by a set of phi parameters. These phi parameters are then learned such that the variational distribution $q_\theta(W)$approximates the exact posterior $p(W/X)$ as closely as possible. The standard measure of closeness between probability distributions is the Kullback-Leibler divergence (KL-divergence) (Hoffman et al., 2013), which leverages Shannon's information theory (Shannon, 1948). The KL-divergence calculates the average number of additional bits that would be required to encode a sample from P using a code optimized for q. In Bayesian inference, the KL-divergence is computed as:

$$D_{KL}(q_\theta(W)||p(W/X) = \int_\emptyset q_\theta(W) log \frac{q_\theta(W)}{p(W/X)} \ dW \qquad 3.34$$

However, there exists a notable issue in this context, namely the requirement to compute $P(W|X)$ in order to calculate $D_{KL}(q_\theta(W)||p(W/X)$. To circumvent this, a distinct formula known as the evidence lower bound ($ELBO$) can be employed as a loss function, which is straightforward to derive.

$$\int_\emptyset q_\theta(W) log \left(\frac{p(W,X)}{q_\theta(W)}\right) \ dW = \log(p(X) - D_{KL}(q_\theta(W)||p(W/X) \qquad 3.35$$

Due to the fact that $log(P(X))$ is exclusively determined by the prior, the minimization of $D_{KL}(q_\theta(W)||p(W/X)$ is equal to the maximization of the evidence lower bound ($ELBO$).

Given their size, it is crucial to select an optimization technique that is computationally efficient when updating the parameters of neural networks. This is why the gradient descent algorithm with backpropagation is a popular method for training neural networks, as it enables efficient parameter updates. A significant portion of contemporary research in this field is focused on finding novel solutions to address this issue, with Bayes by Backprop Blundell et al. (2015), the local reparameterization trick (Kingma et al., 2015), and flip out (Wen et al., 2018) being among the most prominent approaches.

Bayes by backpropagation

Variational inference is a useful mathematical tool for Bayesian inference, however, it requires modifications to be suitable for deep learning. The primary issue arises from the fact that stochasticity prevents backpropagation from functioning at the internal nodes of a network (Buntine, 1994). Various solutions have been suggested to address this problem, such as probabilistic backpropagation (Hernández-Lobato & Adams, 2015) or Bayes-by-backprop (Blundell et al., 2015). The latter option may be more familiar to those working with deep learning, and therefore, we will focus on Bayes-by-backprop in this tutorial. Bayes-by-backprop is a practical implementation of stochastic variational inference (SVI) combined with a reparameterization trick (Kingma & Welling, 2019), which ensures that backpropagation works normally.

To optimize the network using the prior distribution, variational posterior, and training data, the negative evidence lower bound ($ELBO$) is defined as the loss function.

$$F(X,W) = D_{KL}(q_\theta(W)||p(W/X) - E_{q_\theta(W)}\left[Log\left(\frac{X}{W}\right)\right]$$

$$= \int_\emptyset q_\theta(W) log \frac{q_\theta(W)}{p(W)} \; d\theta - \int_\emptyset q_\theta(W) \log p(X/W) \quad d\theta \qquad 3.36$$

As the objective is to minimize $F(X,W)$ using gradient descent, computing the gradients of the two expectations in Equation (3.36) is not feasible analytically, and instead requires Monte Carlo sampling. Blundell et al. (2015) proposes a modification of the local reparameterization trick (Kingma et al., 2015) to obtain unbiased gradient estimates. The derivative of an expectation can be expressed as the expectation of the derivative under certain conditions. For a function $f(w;\ \theta)$, a random variable E with distribution $q(E)$, and a $w = t(\theta;\ E)$ with marginal distribution such that $q(w)dw = q(E)dE$, the following equation is valid:

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial\theta}} E_{q_\theta(W)}[f(W,\theta)] = E_{q(\varepsilon)}\left[\frac{\partial f(W,\theta)}{\partial W}\frac{\partial W}{\partial \theta} + \frac{\partial f(W,\theta)}{\partial \theta}\right] \qquad 3.37$$

The technique of using $f(w_i;\ \theta) = \ log\ q(w_i) - \ log\ p(w_i)\ p(X|w_i)$, where w_i is the ith Monte Carlo sample from the variational posterior $q(w)$, allows for the approximation of the gradients of the loss function 5.1 through Monte Carlo sampling. The number of samples,

M, needed is often low, with $M = 1$ producing sufficient results. In Gaussian variational posterior, the goal is to find the mean, m, and variance, ρ^2, of the variational posterior, assuming a diagonal covariance matrix. Blundell et al. (2015) optimization procedure can be generalized to allow for an arbitrary M. The variables are initialized to set the variational posterior to a standard normal distribution, with m as 0 and p as 1 as per appendix A

Algorithm (3): Bayes by Backprop

```
$Set\ \emptyset = \emptyset_0$
$\boldsymbol{for\ t = 1\ to\ N}$
    $sample\ \epsilon \sim q(\epsilon)$
    $W = t(\epsilon, \emptyset)$
    $f(W, \emptyset) = \log\left(q_{\emptyset}(\mathrm{W})\right) - \log(p(y/X, W)P(W)$
    $\Delta_{\emptyset} f = backprop_{\emptyset}(f)$
    $\emptyset = \emptyset - \alpha\Delta_{\emptyset} f$
$\boldsymbol{End\ for}$
```

Other Machine Learning Methods

K-Nearest Neighbour (kNN)

The purpose of the classification algorithm is to allocate anonymous data to the class containing the most comparable labelled samples. For both the training and test datasets, observational parameters are obtained. While using the kNN method, the datasets must be readied. After predicting the outcome with the kNN method, the model's analytical accuracy should be evaluated. The most commonly used statistic to represent the kNN algorithm is average accuracy. The k value, distance computation, and selection of acceptable predictors all substantially impact model performance. By default, the kNN function uses Euclidean distance, which the following equation can specify. (Hastie et al., 2009).

$$Dis(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \qquad 3.38$$

Where p and q are the subjects to be evaluated with n features, other techniques of calculating distance exist, such as the Manhattan distance. Another notion is the k parameter, which determines how many neighbors are chosen by the kNN algorithm. The selection of k considerably impacts the diagnostic performance of the kNN algorithm. A higher k decreases the influence of variability produced by random error, but it runs the risk of disregarding minor but significant patterns. The key to selecting a fair k value is to find a happy medium between errors and regression problems.

SVM (Support Vector Machines)

Support Vector Machines (SVMs) classification method is proposed for cluster analysis without prior knowledge of input classes. The algorithm begins by running a bipolar classification model on randomly labeled vectors until initial convergence is achieved. At this stage, SVM confidence values for each training instance are obtained. The data with the lowest confidence (worst mislabeled data) is then relabeled with the other class label. The SVM is re-run on the dataset with the partially relabeled data, benefiting from the previous convergence and reduced misidentification penalties. This approach mitigates the risk of singularity traps observed in other methods. By retraining the SVM after each relabeling on the worst misclassified vectors, specifically those with confidence factor values above a certain threshold, the approach addresses its poorly convergent outcome (Schölkopf et al., 2002).

Random Forest:

Random Forest, as explored by Louppe (2014), is an ensemble method in machine learning. It is a versatile approach for classification and regression that leverages the strength of multiple decision trees to create a robust and precise model. The fundamental concept of Random Forest involves constructing a forest of decision trees, with each tree built using a random subset of the training data and a random subset of features. This randomization strategy serves to mitigate overfitting issues and enhances the model's capacity for generalization. The trees in the forest are trained independently and their predictions are aggregated through a majority vote to yield the final prediction.

Random Forest offers numerous advantages compared to alternative machine learning algorithms. Firstly, it is a non-parametric method, eliminating the need for assumptions about data distribution. This flexibility enables Random Forest to accommodate various types of data effectively. Secondly, it exhibits high scalability, enabling it to handle extensive datasets with complex, high-dimensional feature spaces. This scalability makes Random Forest suitable for large-scale applications. Furthermore, Random Forest provides a valuable feature importance measure, allowing users to identify the most significant features for the classification task at hand. This information aids in understanding the underlying factors that drive the model's predictions and supports feature selection processes (Kursa & Rudnicki, 2010).

Random Forest is a versatile machine learning algorithm applicable to both classification and regression tasks. Its foundation lies in constructing an ensemble of decision trees using various subsets of the input data. By combining the predictions of these individual trees, Random Forest enhances the overall accuracy of the model. This ensemble approach empowers Random Forest to effectively handle complex data patterns and deliver reliable results across diverse problem domains.

The Random Forest algorithm excels in handling intricate and nonlinear associations between input features and target variables. By introducing randomness during the model construction, it effectively mitigates overfitting. Unlike a single decision tree, a random forest combines predictions from multiple trees and determines the final outcome based on the majority vote. Increasing the number of trees in the forest not only improves accuracy but also reduces the risk of overfitting, thereby ensuring a more reliable and robust model.

Naïve Bayesian

Naive Bayes is a popular probabilistic machine learning algorithm utilized for classification tasks. It leverages Bayes' theorem, which enables the calculation of the probability of a hypothesis given observed evidence. The "naive" aspect of Naive Bayes stems from its simplifying assumption of feature independence, although this may not hold true in real-world data. Nonetheless, Naive Bayes remains a powerful algorithm capable of achieving high accuracy in various classification tasks (Brownlee, 2019). It constructs a model of the probabilities of each input feature for every possible output class. When presented with a new input instance, the algorithm computes the probabilities of it belonging to each potential class based on the model. It then selects the class with the highest probability as the predicted output.

One significant advantage of Naive Bayes is its simplicity and speed. It operates effectively even with limited training data, swiftly generating accurate predictions for large datasets. Furthermore, Naive Bayes can handle high-dimensional feature spaces and exhibits resilience towards irrelevant features. Its versatility is evident in its successful applications across diverse domains, including text classification, spam filtering, and image recognition. Given its simplicity, efficacy, and broad utility, Naive Bayes remains a favoured choice for numerous machine learning applications (Sammut & Webb, 2011).

Summary

In this chapter, we mainly focus the patient's profile as they had hospitalized due to the being infected with Covid-19. For this purpose, two particular hospitals would be chosen as they are known for having Covid-19 cases from the beginning and one belongs to public sector while the other one is private sector. The dataset ends up with huge number of variables, for instance, age, gender, status (cured/died), lab tests (including blood test, chest test, urine test and etc), oxygen measurements, blood pressure measurements and others. We have no access to sensitive information such as, name of the patients, telephone numbers or any other means to be non-anonymous. This means the dataset is fully anonymous, thus neither consent form is nor ethical paper is needed. Hence, this current study considers on a secondary data analysis of a cross-sectional study type since the information was already collected by the hospitals itself.

Both traditional mathematical and data mining approaches will be used in this study. Traditional statistical methods provide distributions that describe an observable property (descriptive statistics) which are used to determine the reliability of a sample taken from a population (inferential statistics). They are focused on continuously measuring the properties of objects with the goal of predicting the frequency with which such effects will occur when the measuring operation is replicated at random or stochastically, and the established hypotheses are then tested against the evidence (Bzdok et al., 2017).

CHAPTER IV

Statistical Analysis Results

Descriptive and Visualization Analysis

It is begun with descriptive statistics since it is considered as a vital stage of any study's outcomes where initial thoughts on the nature of the dataset are noticed as well as sometime enables the researchers to find potential patterns among the explanatories and response variable. To simplify the discussion and easy to follow, it was split to two parts, exploring the association of predicted variable with quantitative and qualitative covariates separately.

Response Association with Quantitative Variable

Table 3 catches attention on the effect of the laboratory measurements on the response variables and whether the increase or decrease unit of any of them caused to Covid-19 patients to death. According to the outcomes, it was obvious that except Netrophil variable, there were highly significant differences between mean values of the survivors and died cases with p-values <0.001. For instance, mean value of HR for those discharged alive in the hospital was measured with (84.78 ± 13.80) while it was (107.92 ± 16.46) for those who died, thus it led us to report that the disease had impact on increasing heart rate pulse. Furthermore, also great difference of mean values was noticed for CRP variable with 16.60 and 48.37 for recovered and died cases respectively. When compared to upper respiratory tract infection, those dies with Covid-19 were more likely to be diagnosed with pneumonia mean value (22.69 ± 4.6901 vs 18.81 ± 1.85, p-value<0.001), and mean value of quadrant (2.51 ± 1.07 vs. 3.183± 1.13, p-value<0.001) and pulmonary (33.28 ± 17.87 vs. 39.76 ± 16.93, P-value<0.001) were slightly higher and significantly differed. As stated, no significant difference was occurred due to Neutrophil results between died and recovered cases with (3.1601 ± 0.4974 vs. 3.1403 ± 0.68, p-value = 0.70). On admission, patients with severe/critically ill Covid-19 caused to death had higher temperatures, lower SpO2, and higher CT image quadrant scores and pulmonary opacity values.

Table 3:

Descriptive Statistics of Laboratory Results Under Investigation Recovered and Died Cases

Variables	Recovered				Died				P-Value[a]
	N	Mean ± SD	Median	IQR	N	Mean ± SD	Median	IQR	
Temperature	341	37.56 ± 1.97	38.30	3.90	196	38.30 ± 1.66	38.40	1.78	0.000
HR	341	84.78 ± 13.79	84.60	17.90	196	107.92 ± 16.46	107.40	22.75	0.000
Respiratory	341	18.81 ± 1.85	18.70	2.50	196	22.69 ± 4.69	21.90	6.60	0.000
Quadrant	341	2.52 ± 1.087	2.00	1.00	196	3.18 ± 1.13	4.00	1.00	0.000
Pulmonary	341	33.28 ± 17.87	30.00	25.00	196	39.76 ± 16.93	37.00	35.00	0.000
Neutrophil	341	3.16 ± 0.50	3.10	0.95	196	3.14 ± 0.68	3.10	1.00	0.699
Lymphocyte	341	1.37 ± 0.27	1.40	0.50	196	0.78 ± 0.21	0.75	0.30	0.000
Platelet	341	192.91 ± 22.97	200.10	21.00	196	172.70 ± 27.02	178.10	47.00	0.000
Albumin	341	41.49 ± 2.27	41.60	3.70	196	38.55 ± 2.23	38.60	2.90	0.000
Creatinine	341	66.63 ± 7.42	68.30	13.95	196	68.28 ± 9.11	68.10	13.90	0.024
APTHT	341	33.53 ± 2.55	33.50	4.00	196	32.28 ± 2.28	32.70	4.90	0.000
Fibrinogen	341	3.3 ± 0.40	3.30	0.70	196	4.45 ± 0.83	4.20	1.40	0.000
SpO2	341	96.68 ± 3.65	98.20	4.85	196	91.18 ± 4.40	91.20	5.30	0.000
WBC_Count	341	4.96 ± 0.93	5.20	1.50	196	3.84 ± 0.76	3.70	0.50	0.000
CRP	341	16.60 ± 19.99	8.30	12.65	196	48.37 ± 30.81	51.20	54.60	0.000
D_dimer	341	0.45 ± 0.23	0.40	0.30	196	0.62 ± 0.24	0.65	0.40	0.000

[a] Continuous variables: T-test or Mann-Whitney tests as appropriate

In addition, infected cases who died from Covid-19 had higher C-reactive protein, fibrinogen, and D-dimer levels than recovered cases. Similarly, those who passed away had lower APTHT, lymphocyte, platelet, and albumin counts as per the results provided in Table 3.

According to box-plots shown in Figure 6, the differences and the distributions of the variables were clearly provided. Normality assumption can also be seen were the boxes inside the plots for almost all of them had normal shapes with very small amounts of outliers.

Figure 6:

Box-Plot Illustration for continuous covariates

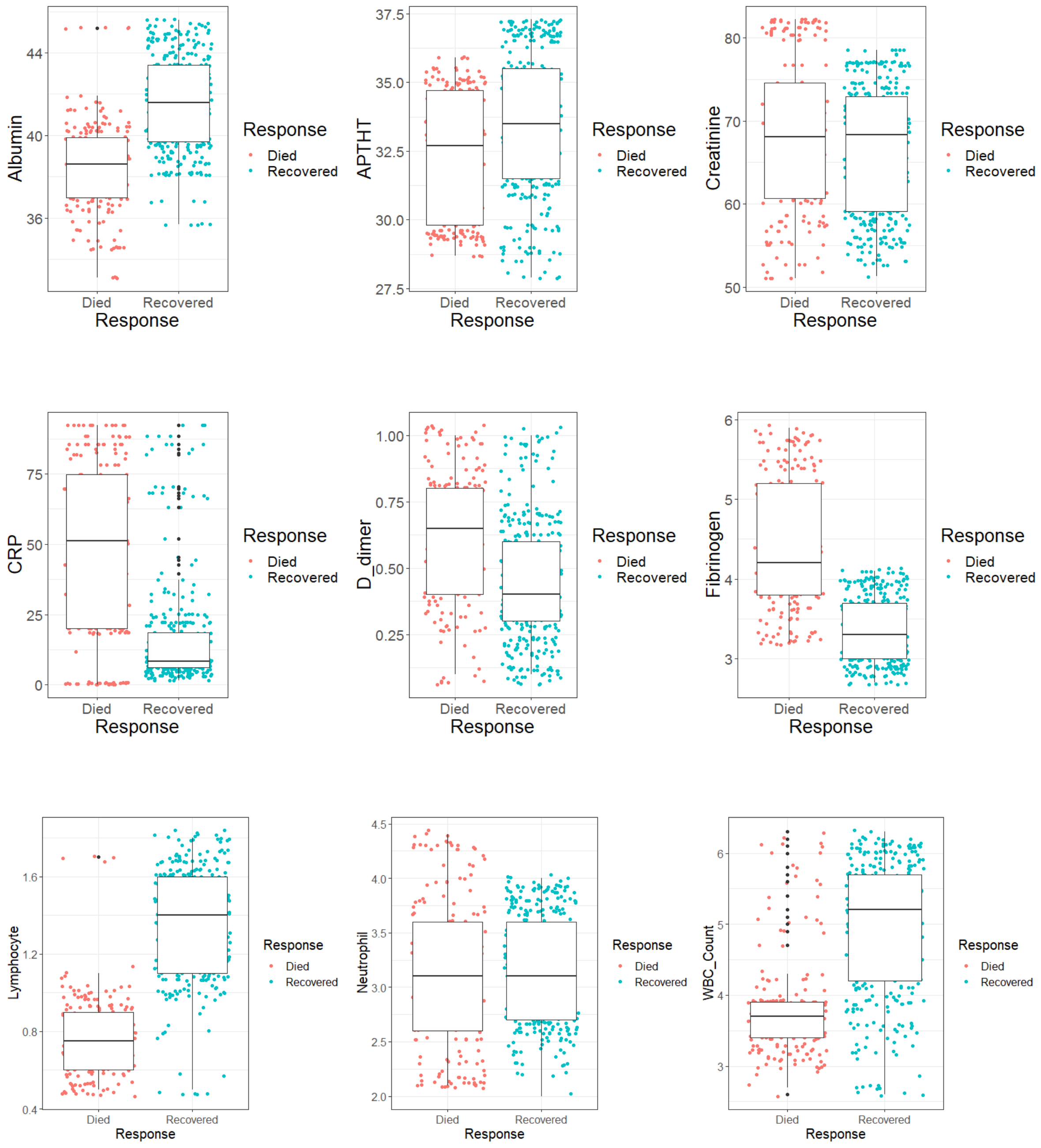


Figure 7 illustrates the distribution of the variables against the response variable where one can simply identify the effectiveness of the covariates on dependent variable. For example, regarding CRP, HR, SpO2, Lymphocyte and WBC_Count measurements, the distribution of the survived and died cases were clearly split for two different areas with a very low rate of overlaps, and these variables were already providing insights to be listed on predicting the probability not surviving a patient. However, it was not wise to decide at this stage to highlight variables had impact on increasing the risk of dying from the disease since further tests are required to be implemented.

Likewise, we can detect that Pulmonary, Quadrant, Temperature, Neutrophil and Creatinine quantities were not highly associated with the response variable based on their distributions shown in Appendix D.

Response Association with Qualitative Variables

The study calculated figures for nine categorical variables were counted to predict the probability of death due to Covid-19. Table 4 calculates the ratio in death rates per gender (male death rate: female death rate) in Covid-19 patients. The male death rate was 25% times higher than the overall female death rate. This means that according to the cohort included in this study, men make up to 20.5% of all Covid-19 deaths while only 16% of the death rate was recorded as women.

With respect to Smoking factor, a total of 537 Covid-19 patients are included in our analysis, 196 of whom (36.5%) experienced disease progression and 214 (40%) with a history of smoking. Among those with a history of smoking, 13.4% experienced disease progression and died, compared with 23.1% of non-smokers. The analysis showed non-significant association between ever smoking and Covid-19 progression. Moreover, the results of this statistic analysis demonstrated that stroke was not significantly associated with Covid-19 mortality as shown in the below table. The mortality rate among patients who had stroke was lower than those without stroke that is 13.8% and 22.7% respectively. On the contrary, patients with no previous history of stroke had shown greater improvement and recovery which is 38.6% compared to recovery among patients who had previous history of stroke which is 25%.

Figure 7:

Density Distribution with respect of response variable

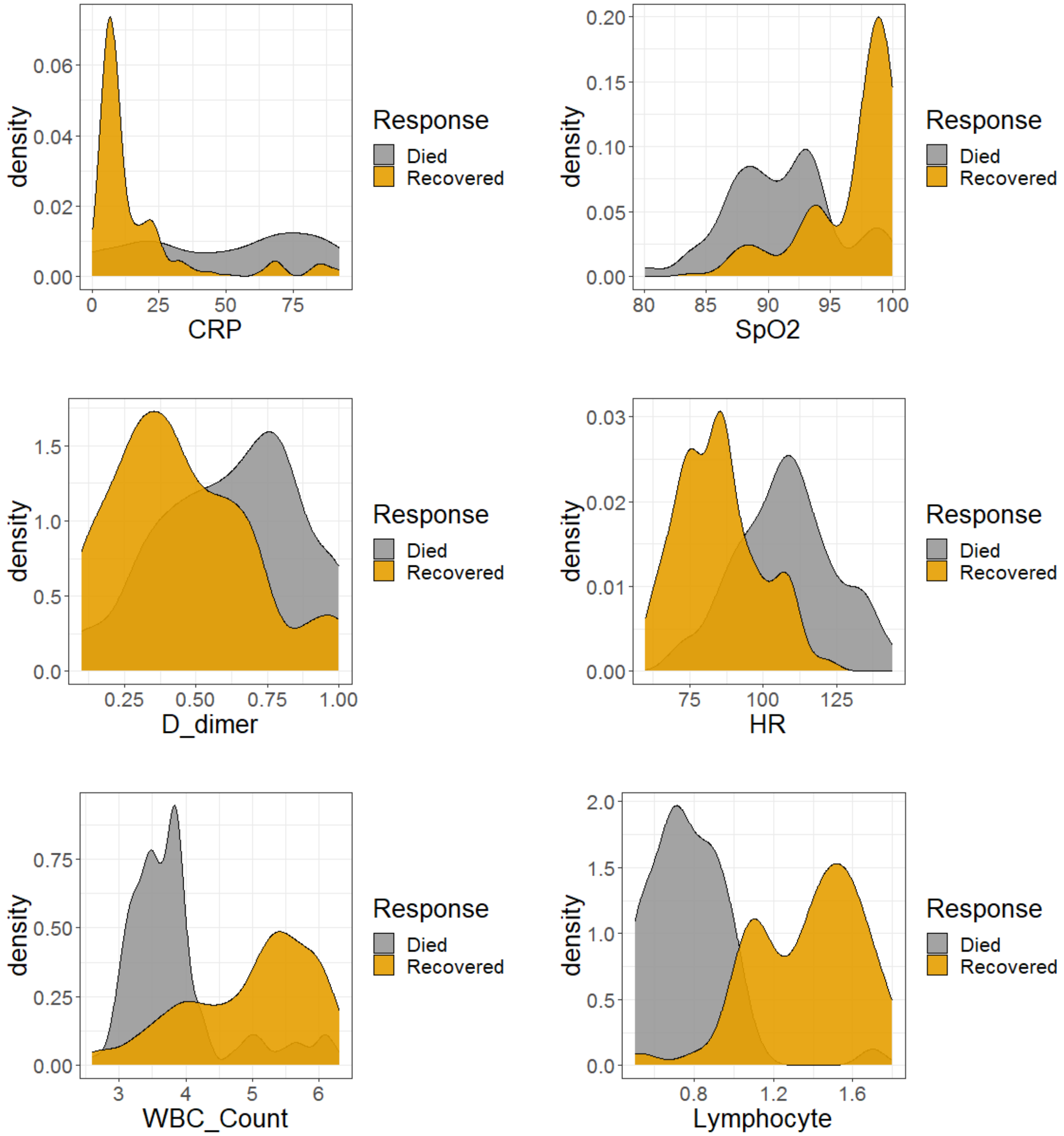


Table 4:

Descriptive Statistics of Categorical Variables Associated with Response Variable

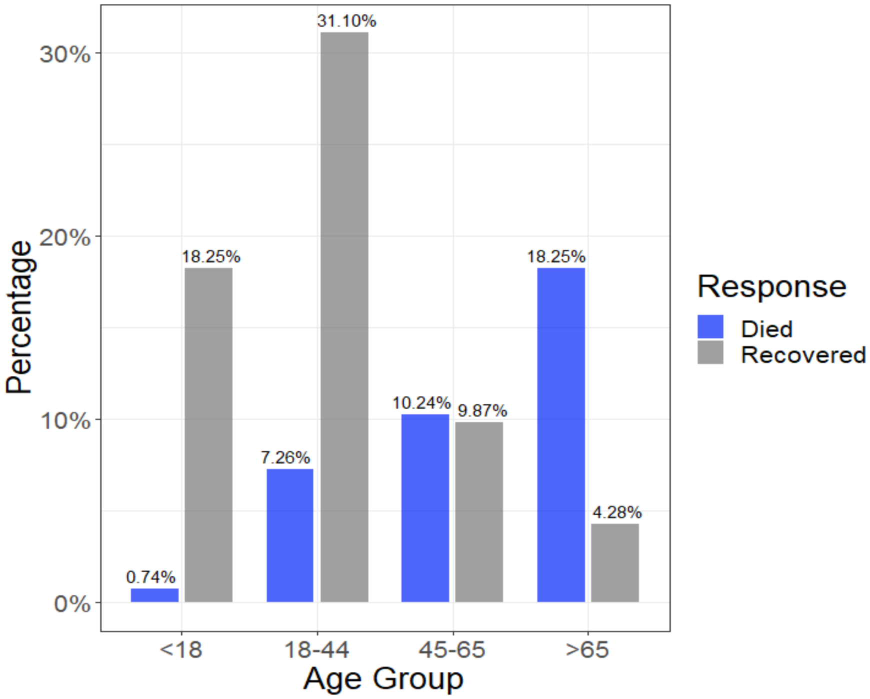
Variables	Levels	Recovered		Died		P-Values^{b}
		N	%	N	%	
Gender	Male	199	37.1%	110	20.5%	0.614
	Female	142	26.4%	186	16.0%	
Age	<18	98	18.2%	4	0.7%	0.000
	18-44	167	31.1%	39	7.3%	
	45-64	53	9.9%	55	10.2%	
	65+	23	4.3%	98	18.2%	
Smoking	Yes	142	26.4%	72	13.4%	0.263
	No	199	37.1%	124	23.1%	
Fever	Yes	137	25.5%	112	20.9%	0.000
	No	204	38.0%	84	15.6%	
Cough	Yes	142	26.4%	126	23.5%	0.000
	No	199	37.1%	70	13.0%	
Sputum	Yes	99	18.4%	75	14.0%	0.028
	No	242	45.1%	121	22.5%	
Hypertension	Yes	135	25.1%	133	24.8%	0.000
	No	206	38.4%	63	11.7%	
Diabetes	Yes	119	22.2%	133	24.8%	0.000
	No	222	41.3%	63	11.7%	
Stroke	Yes	134	25.0%	74	13.8%	0.724
	No	207	38.5%	122	22.7%	

[a]Categorical variables: Fisher Exact or Chi-square tests as appropriate

Age variable has been one of the main factors of mortality rate among Covid-19 patients as shown in Figure 8 and test analysis in Table 4. 18.25% of death rate was among patients aged over 65 years. The (18-44) age-group was the most affected by a wide margin meanwhile the recovery rate was at the peak in this age group which is 31.10%. Fatality rate among those under ages of 18 was only 0.74% which is considerably low compared to the rest of the other age groups.

Figure 8:

Bar chart presentation of Age with respect to Response variable levels

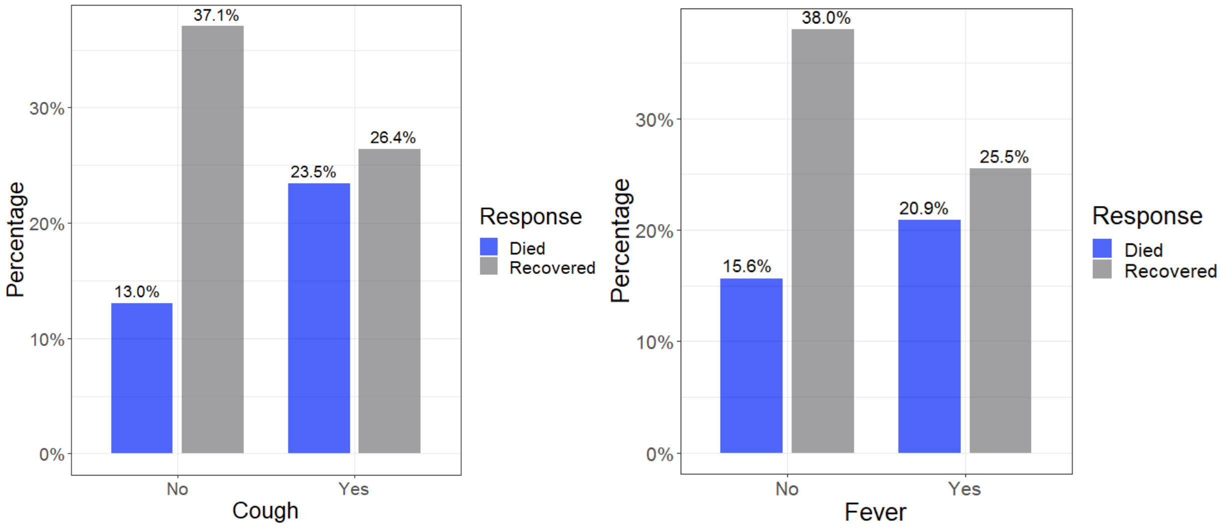


Furthermore, as shown in Figure 9, the cohort included in this study showed that 23.5% of the patients died had sever coughing whereas only 13% of the death rate was among those without coughing per Figure 9.A. This explains that acute respiratory distress syndrome was correlated with mortality rate. In addition, Fever has been one of the main factors that is common to the majority of hospitalized Covid-19 patients. As presented in Figure 9.B, a high body temperature corelates with the mortality in Covid-19 patients to a great extent. Patients with body fever $\leq 36\,°C$ had significantly higher mortality compared to normothermia patients.

In addition, patients were suffering from diabetes revealed significant death rate compared to non-diabetes holders as illustrated in Figure (10.a). The fatality among patients with diabetes was more than double with those without diabetes, 24.8% and 11.7%

respectively. It is worth noting that the recovery among patients without diabetes was significantly higher compared to the ones with diabetes that is 41.3% and 22.2% respectively. Also, the mortality rate in patients with Covid-19 was reviewed.

Figure 9:



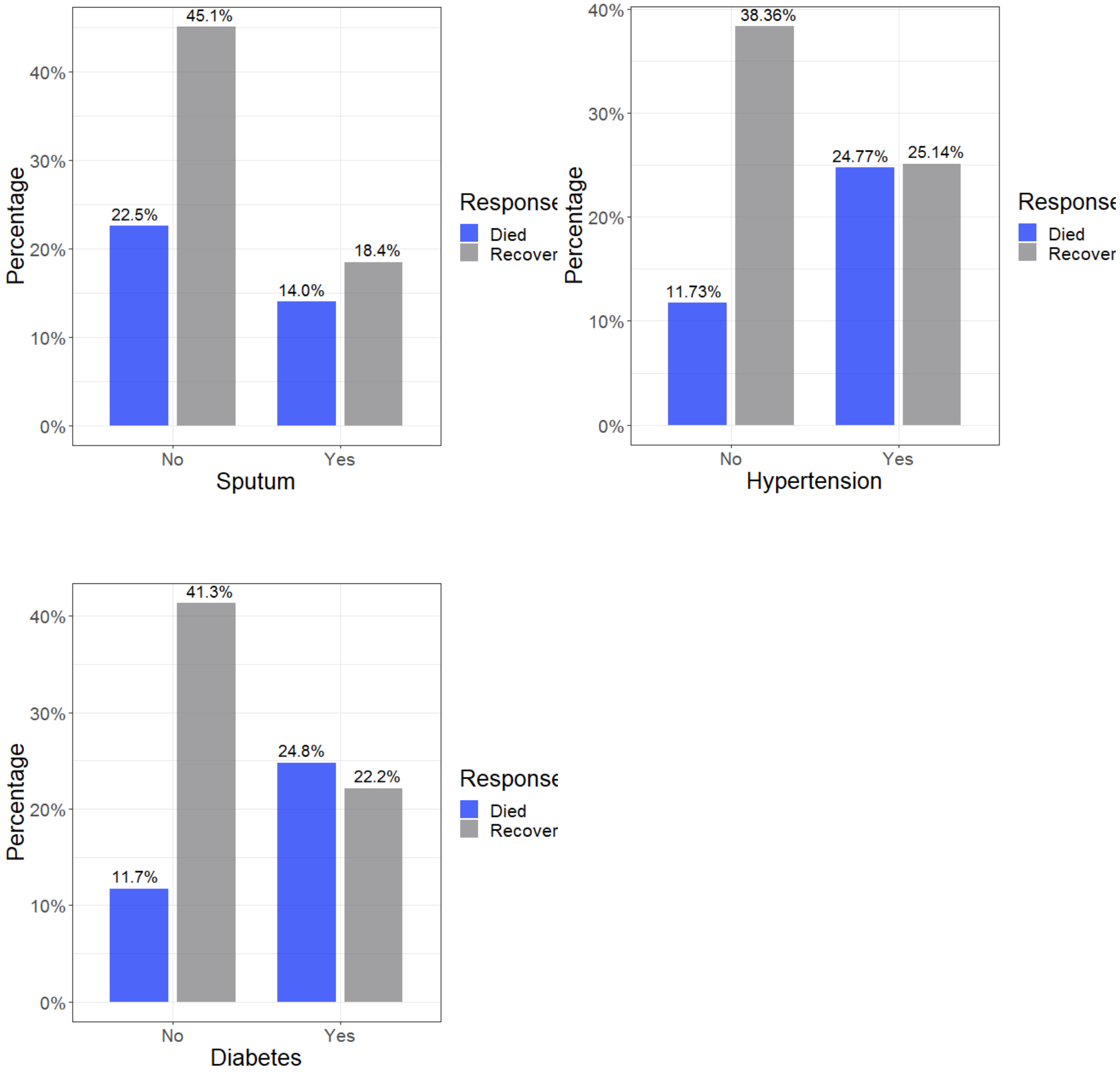
A) Cough vs Response Bar plot

B) Fever vs Response bar plot

The analysis showed that the overall mortality rate was 24.77% among those patients with hypertension. Hypertension assessment results showed high death rates in the results of this study as presented in Figure (10.b). Of the 537 patients sampled, sputum is counted for 14% of the deaths whereas death rate among patients without sputum was much higher which is 22.5% as shown in Figure (10.c). The recovery among patients with no sputum was significantly higher 45.1% compared to those with sputum 18.4%.

Figure 10:

Bar charts for A) Diabetes, B) Hypertension, C) Sputum

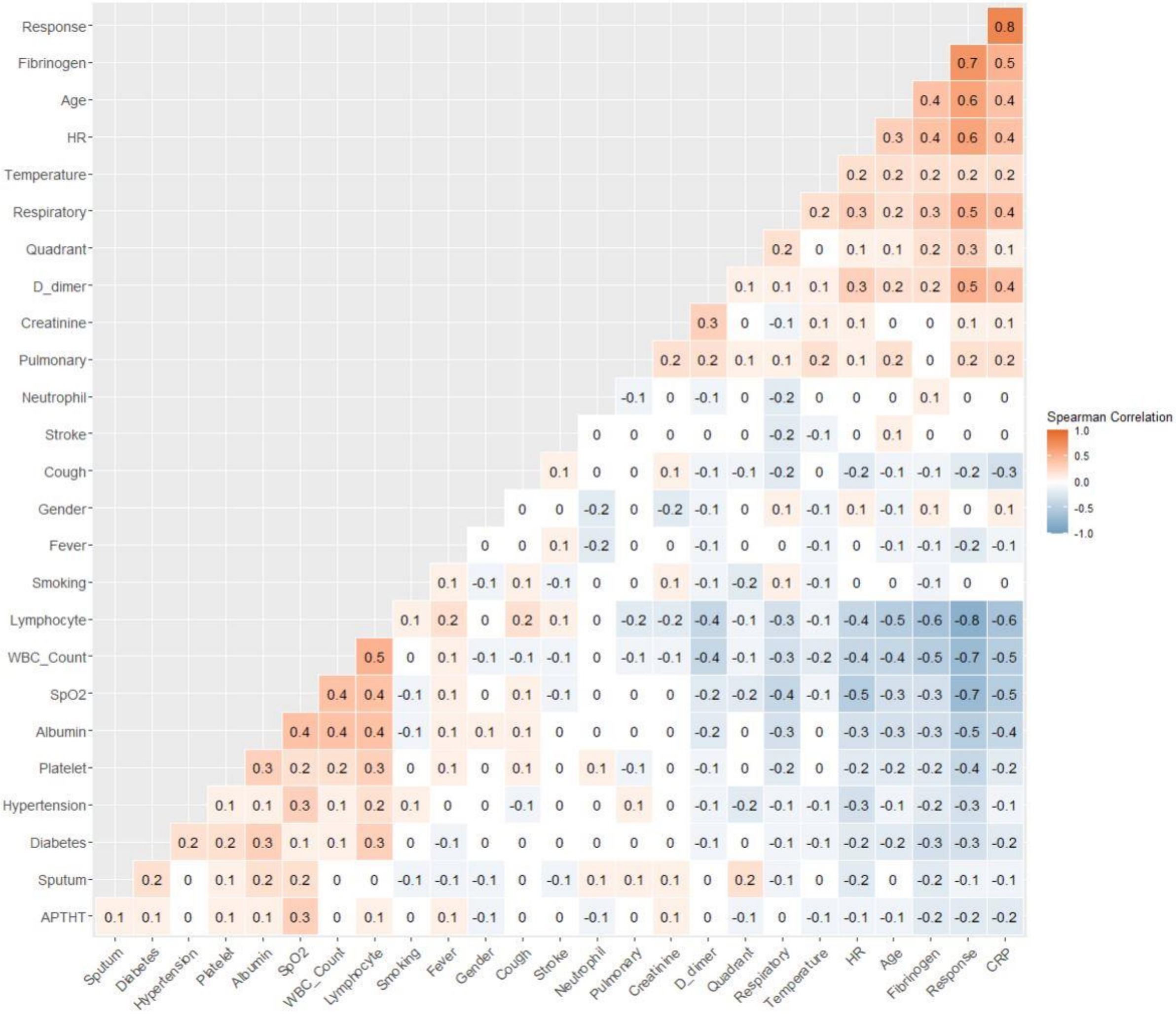


Model Building Analysis:

To avoid multicollinearity issues which results high p-value as well as unreliable estimated parameters while building up models with especially multiple variables which of course influence the predicted values afterwards, and it was most likely believed there might be high correlation among the covariates. Hence, correlation matrix was plotted in order to easily detect how the variables highly linked to each other and referencing to Figure 12, it was noted that Fibrinogen had moderate and positive relationship with Age and HR, Age and Lymphocyte with 0.4,0.4 and -0.6, while Lymphocyte was found to correlated with WBC_count and CRP with 0.5 and -0.6 respectively. Consequently, among any related pairs, only one retained to be included in the model selection process and 16 of the 26 independent variables were retained for the next step.

Figure 11:

Correlation Heatmap Matrix For All Variables Study



Univariate Model:

We start with fitting classic simple Logistic regression as well as Bayesian Logistic regression on the trained dataset where the spilt was made based on 70% for training the model and 30% for the testing model, all the remaining explanatory variables from previous section were inputted each at a time and their AIC along with Nagelkerke R-Sqaure values were measured and tracked to evaluate the changes. Table 5 demonstrates the univariate outcomes for both approaches and although no considerable differences were found in terms of the coefficient values, the changes from their SE values were important to report. Table 5 explores the SE values of the coefficients estimated from MCMC had lower SE in most of the models since the lower SE, the better and more reliable value is. As a result, this helped us to choose Bayesian Logistic over classic Logistic regression.

Table 5:

Univariate Logistic Regression Analysis

Models	Vairables	MLE Approach			MCMC Approach		
		Coefficients	SE	P-value	Coefficients	SD	P-value
Model 1	Age (18-44)	2.0130	0.7533	0.0075	1.9170	0.7050	0.0167
	Age (45-65)	3.6306	0.7578	0.0000	3.5540	0.7029	0.0175
	Age (>65)	5.1066	0.7751	<0.001	5.0490	0.7217	0.0180
Model 2	APTHT	-0.1831	0.0466	<0.001	-0.1699	0.0437	<0.001
Model 3	Cough	1.0477	0.2316	<0.001	-1.0447	0.2297	<0.001
Model 4	CRP	0.0430	0.0049	<0.001	0.0435	0.0049	<0.001
Model 5	D_dimer	3.2779	0.5304	<0.001	3.2600	0.5211	0.0036
Model 6	Diabetes	1.0596	0.2306	<0.001	-1.0600	0.2267	<0.001
Model 7	Fever	0.5621	0.2247	0.0120	-0.5635	0.2225	0.0015
Model 8	Hypertension	1.0244	0.2327	<0.001	-1.0248	0.2288	<0.001
Model 9	Platelet	-0.0300	0.0047	<0.001	-0.0295	0.0045	<0.001
Model 10	Pulmonary	0.0176	0.0063	<0.001	0.0176	0.0062	<0.001
Model 11	Quadrant	0.5589	0.1079	<0.001	0.5630	0.1067	<0.001
Model 12	SpO2	-0.2912	0.0326	<0.001	-0.2943	0.0324	<0.001
Model 13	Sputum	0.4770	0.2340	0.0410	-0.4767	0.2290	0.0560
Model 14	Stroke	0.0903	0.2280	0.6920	-0.0901	0.2237	0.6430
Model 15	Temperature	0.2267	0.0637	<0.001	0.1829	0.0550	<0.001
Model 16	WBC_Count	-1.3782	0.1573	<0.001	-1.3690	0.1523	<0.001

Another way to evaluate the models is to compute AIC as well as Nagelkerke R-Sqaure values for each model separately and the lower AIC is the better model however the higher R-square is the preferrable model. According to Table 6, AIC extracted from Bayesian models had lower values than that from classic Logistic regression models whereas their R-square values were quite identical. It was worth mentioning that model 1 where the effect of Age variable entered had the lowest AIC with 314.500 as well as had 48% total variability on predicting death cases, followed by SpO2 refers to patient's oxygen with 37.8% percentage of changes on the probability of dying someone where his/her oxygen level was not stable. White blood cell count was found to be as effective as SpO2 with producing changes on probability of dying a Covid-19 contacted case with 37.5%. Furthermore, CRP was also discovered to be a significant factor in infected individuals to Covid-19 disease, accounting for 35.4% and there were likewise other effective variables such as Platelet and D-dimer where had relatively high R-Square values with 17.00% and 16.30% respectively. Their associations can easily be noticed in Figure 13.

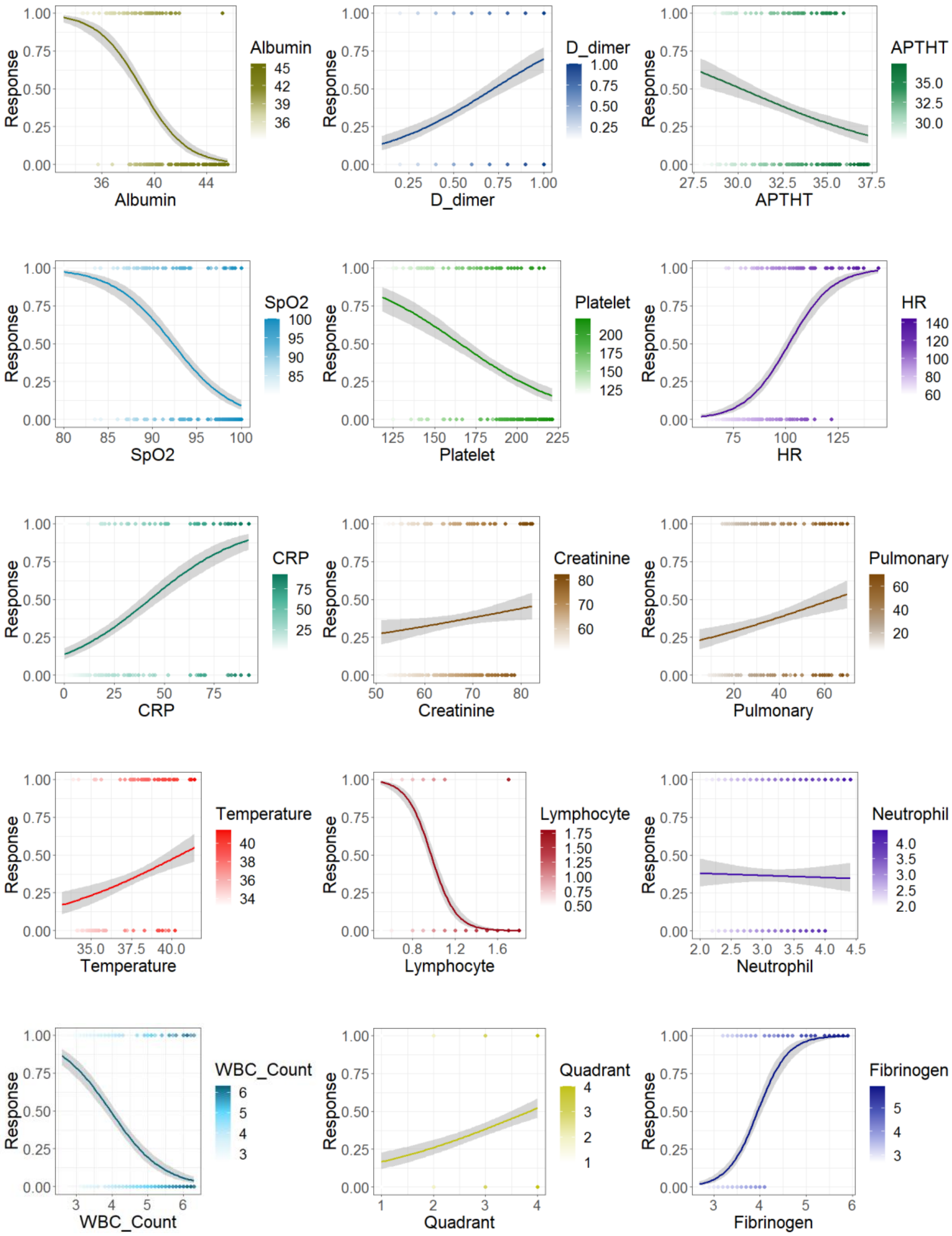
Table 6:

Diagnosis Test Result For Univariate Logistic Regression Models (MLE And MCMC) Parameter Estimation

Models	MLE Approach		MCMC Approach	
	AIC	Nagelkerke R-Sqaure	AIC	Nagelkerke R-Sqaure
Model 1	314.500	48.00%	302.6572	47.96%
Model 2	444.5900	6.20%	442.5915	6.20%
Model 3	439.4300	8.10%	437.4257	8.12%
Model 4	356.5500	35.40%	354.5629	35.39%
Model 5	416.5800	16.30%	414.5797	16.30%
Model 6	438.8400	8.30%	436.8356	8.34%
Model 7	454.4100	2.50%	452.4143	2.46%
Model 8	440.4900	7.70%	438.4882	7.73%
Model 9	414.5900	17.00%	412.6027	16.98%
Model 10	452.8500	3.10%	450.8506	3.06%
Model 11	431.0600	11.20%	429.0620	11.18%
Model 12	348.2100	37.80%	346.2192	37.80%
Model 13	456.5800	1.60%	454.5760	1.62%
Model 14	460.5600	0.10%	458.5650	0.06%
Model 15	447.1100	5.20%	445.5916	5.07%
Model 16	349.2000	37.50%	347.2044	37.52%

Figure 12:

Association of Quantitative variables with response using Logistic Regression (MCMC)



In classification modeling, performance evaluation holds great significance, and the AUC-ROC Curve serves as a valuable tool for assessing the fitted model. Classification accuracy, which measures the proportion of correctly classified cases by a classifier model out of the total cases, is a fundamental metric for evaluating model performance, particularly for unbalanced data. Sensitivity, specificity, precision, recall, and the AUC-ROC curve are additional commonly used performance measures closely related to classification accuracy.

The AUC-ROC curve, also known as the Area Under the Receiver Operating Characteristics curve, is a vital evaluation metric for gauging the performance of any classification model. It quantifies the degree of separability, with AUC representing the measure of separability and ROC representing the probability curve. The AUC reflects the model's ability to correctly predict class 0 as 0 and class 1 as 1. For instance, in the context of Covid-19 disease, a higher AUC indicates the model's proficiency in differentiating between patients who have recovered and those who have succumbed to the illness.

Consequently, the ROC curve was specifically plotted for Bayesian Logistic regression, as this approach exhibited superior parameter estimation compared to the classic approach. By utilizing the AUC-ROC curve, we can effectively evaluate the performance of the Bayesian Logistic regression model in distinguishing between the classes of interest.

Figure 13:

ROC Evaluation Curves for applied univariate models in Bayesian logistic regression

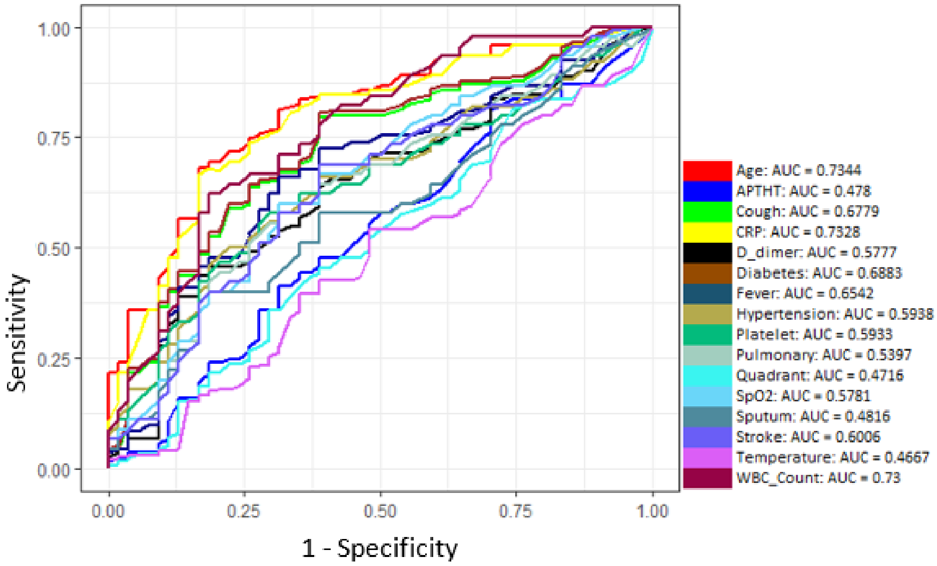


Figure 13 serves us finding the most important factor which was significantly contributed in increasing probability of dying due to Covid-19 disease, and it can be noticed that Age variable at almost all joint points between calculated sensitivity and 1-specificity had highest coordination and was above all the others. According to Table 7, the highest AUC values were calculated for Age, CRP and white blood cells variable with 73.44% and 95% CI (0.729, 0.744), 73.28% with 95% CI (0.728, 0.741) and 73% with 95% CI (0.712, 0.763)) respectively. Followed by diabetes, cough and fever with AUC values 68.83%, 67.79% and 65.42% correspondingly.

Table 7:

Accuracy Analysis for Univariate Logistic Regression models (MLE and MCMC) parameter estimation

Predictors	Accuracy	95% CI	P-Value
Age	73.44%	(0.729, 0.744)	0.0210
APTHT	47.80%	(0.463, 0.481)	0.0000
Cough	67.79%	(0.671, 0.680)	0.0330
CRP	73.28%	(0.728, 0.741)	0.0130
D_dimer	57.77%	(0.562, 0.625)	0.0000
Diabetes	68.83%	(0.677, 0.689)	0.0000
Fever	65.42%	(0.649, 0.660)	0.0000
Hypertension	59.38%	(0.591, 0.601)	0.0254
Platelet	59.33%	(0.512, 0.664)	0.0342
Pulmonary	53.97%	(0.492, 0.632)	0.0012
Quadrant	47.16%	(0.471, 0.476)	0.0037
SpO2	57.81%	(0.572, 0.580)	0.0021
Sputum	48.16%	(0.453, 0.491)	0.0560
Stroke	60.06%	(0.592, 0.624)	0.0674
Temperature	46.67%	(0.452, 0.563)	0.5210
WBC_Count	73.00%	(0.712, 0.763)	0.0341

Building Multivariate Models: With More Than One Explanatory Variables

Forward selection was used to determine which variables should be included in the final model, starting with a simple null assumption. First, the null model with (Status ~ 1) was calculated, and the residual (or "null") deviance was taken. The null model has $(n-1)$ degrees of freedom, where n is the total number of cases of our response variable Status. This model was assumed to be poor, so additional analysis was required. Then, for each explanatory variable, only the response variable was used, yielding Status ~ Age, Status ~

Gender, and so on. The residual deviance was calculated for each model, and the model with the lowest residual deviance was studied further. Each residual deviance has $(n - p)$ degrees of freedom, where p is 2 because we are only interested in one variable in the model plus the intercept, which was also included in the null model. The difference in deviance between our best fitting single variable model and the null model had to be analyzed to see if the new model provided significant improvements. The model with Age variable included was chosen to the next phase and proceeded the same as above (See Appendix 1) where their AIC as well as Nagelkerke R-Sqaure were measured and recorded as shown in Table 8. As a result, this analysis led to the identification of the 7 variables that are most significant.

Table 8:

Result of stepwise forward model selection approach for Bayesian logistic regression (MCMC) and classic logistic regression (MLE)

Models	MLE Approach		MCMC Approach	
	AIC	Nagelkerke R-Sqaure	AIC	Nagelkerke R-Sqaure
Model 1	314.500	48.000%	308.4967	47.970%
Model 2	224.700	68.700%	216.2335	68.660%
Model 3	183.970	76.500%	174.1632	76.430%
Model 4	171.490	78.900%	159.7710	78.870%
Model 5	160.870	81.000%	148.3420	82.310%
Model 6	146.35	83.600%	132.4520	84.643%
Model 7	142.21	84.500%	129.3280	85.212%

Table 9 illustrates the best fitted model from both approaches. Similar to univariate outputs, the same conclusion can be made where Bayesian approach performed better according to the standard errors of the coefficients, although at Age (2) and Age (3) the MLE technique produced lower SE of the parameters.

Table 9:

The best fitted model outcomes by Bayesian logistic regression (MCMC parameter estimation)

Included Variables	Classic Logistic Regression (MLE)			Bayesian Logistic Regression (MCMC)		
	Coefficients	SE	P-value	Coefficients	SE	P-value
Intercept	28.6433	6.1532	0.0000	27.6297	4.2894	0.0000
Age (18-44)	2.2171	0.9668	0.0218	2.4636	1.0315	0.0000
Age (45-65)	4.8459	1.0393	0.0000	4.8139	1.0792	0.0000
Age (>65)	6.0925	1.0975	0.0000	6.0791	1.0708	0.0000
SpO2	-0.3333	0.0655	0.0000	-0.2132	0.0453	0.0000
WBC_Count	-1.1157	0.2649	0.0000	-0.9886	0.2441	0.0000
Diabetes (Yes)	1.2976	0.5109	0.0111	1.0398	0.4518	0.0000
Cough (yes)	2.0109	0.5858	0.0006	1.9824	0.5305	0.0000
Hypertension (Yes)	2.0048	0.5565	0.0003	1.8886	0.4864	0.0000
CRP	0.0225	0.0093	0.0156	0.0268	0.0085	0.0000

Parameter Interpretation for Final Bayesian Logistic Regression:

Table 9 and 10 are the most important part in logistic regression modeling where the magnitude of the variables can be identified. To begin with, Age (18-44) coefficient (2.2171) which stands for 18-44 years old provided in Table 9, is statistically significant (associated with a p-value of 0.05), implying that Age factor does influence risk of being died from Covid19 disease. Because it is a positive number, we can conclude that age raises the risk of developing the disease. Therefore, the odds ratio of Age (18-44) was calculated as approximately 12 with 95%CI (9.380-17.069) as shown in Table 10. This means holding other variables as constant, a patient in age group (18-44) had 12 times higher chance of losing life because of Covid19 compared to individuals who were less than 18 years old, and people in age group (45-65) had 123 more odds to die as well as 436 times more chance in age group more 65 years old than patients were less than 18 years old. It can be noticed that younger people had higher chance to survive from the disease.

Similarly, coughing symptom severely identified among admitted patients had significant impact on increasing the odds to die. This pointed to that inpatient with coughing will rise the odds of being dying by Exp (1.9824) = 7.26 times as seen in Table 10. That being said, inpatient with coughing had 7 times higher chances to die compared to not having strong cough. Like coughing, diabetes was also found to be statistically significant with coefficient (1.2976) and p-value <0.001, and recognized to increase the risk’s odds by Exp (1.2976) =

2.8287. This enables us to report that there was a 182% increase in the odds of passing away with presence diabetes.

In addition, people with possessing covid19 as well as hypertension had high risk and pointing to the result, the factor had positive coefficient value with (2.0048) and statistically significant under 0.05 level were led to increase the logit of predicting dying. Thus, the odds of patient died who had hypertension was 6.6 time higher than patients who did not suffer from hypertension with a 95% CI of (2.495 and 12.097).

Moving to SpO2 which refers to measured oxygen for hospitalized cases, with coefficient (-0.2132), and its odds ratio was (0.80). Hence, SpO2 is associated with a 20% (1 – 0.80 = 0.20) reduction in the relative risk of dying. In addition to that, for a 1-unit increase in the corresponding oxygen's level of patient admitted to hospital due to the disease is associated with a lower risk of dying due to Covid19.

Table 10:

Odds Ration and 95% Confidence Interval of Odds Ratio for Bayesian logistic regression coefficients estimated by MCMC

Variables in the model	Odds Ratio (MCMC Logistic Regression)	95% CI of Odds Ratio	
Age (18-44)	11.7469	9.380	17.069
Age (45-65)	123.2075	106.592	175.452
Age (>65)	436.6317	421.492	521.402
SpO2	0.808	0.630	0.815
WBC_Count	0.3721	0.195	0.551
Diabetes (Yes)	2.8287	1.345	9.963
Cough (Yes)	7.2601	2.370	11.547
Hypertension (Yes)	6.6104	2.495	12.097
CRP	1.0272	1.004	1.042

Referencing to the effect of white cell count in blood on the probability of dying holding other variables as constant, the estimated parameter was (-0.9886), means that one unit increase from white blood cell, the logit of predicting dying deceases. Turning to odds ratio, it was estimated to be 0.37; for every 1 increase in white blood cell count, an infected Covid19 person had 63% chance to survive than not with a 95% CI of (0.195 and 0.551).

The CRP turned out to be contributing in increasing the odds of dying with coefficient (0.0225), which is positive. And this means that a rise in CRP was associated with an increase in the likelihood of being dead due to Covid19. It was important to state that CRP was associated with a 2.72% rising the chance to die. Although the effect size small, it was statistically significant with p-value < 0.001.

MCMC Result Assessment (Posterior Distribution)

Traceplots and Histograms of Posterior Distribution

The resulting posterior distributions and some diagnostic tests were required to evaluate the estimated parameters. The standard procedure is to set the number of MCMC trials (usually several thousand) and discard the first 1000 trials or more as a burn-in. Following that, one of the first tests is usually a traceplot analysis. Consider our results for a 21,000-trial MCMC analysis with a 3000-trial burn-in and every other result pruned. To begin, look at the traceplots for the first 100 trials after the burn-in (see Figure 8). Furthermore, each of the unknown parameters' posterior distributions can be summarized as a histogram, which provides the shape of the parameter's marginal distribution.

The 95% Bayesian credible intervals for each of the three parameters were calculated although there are various approaches to this. The 0.025 and 0.975 quantiles for each parameter are shown in Table 11. (along add in the 25th, 50th, and 75th percentiles for fun). The credible intervals for each parameter can be found by listing the values in the first and last rows. The credible interval for the Age appears to be positive and does not include zero, implying that there is evidence that Age group was positively related to odds of not surviving from the disease. However, SpO2 as well as WBC were negatively associated the response variable confirming negative link. The credible intervals also provide us insights about the posterior distribution whether can be reliable or not.

Table 11:

Credible Interval Finding of Posterior Parameters

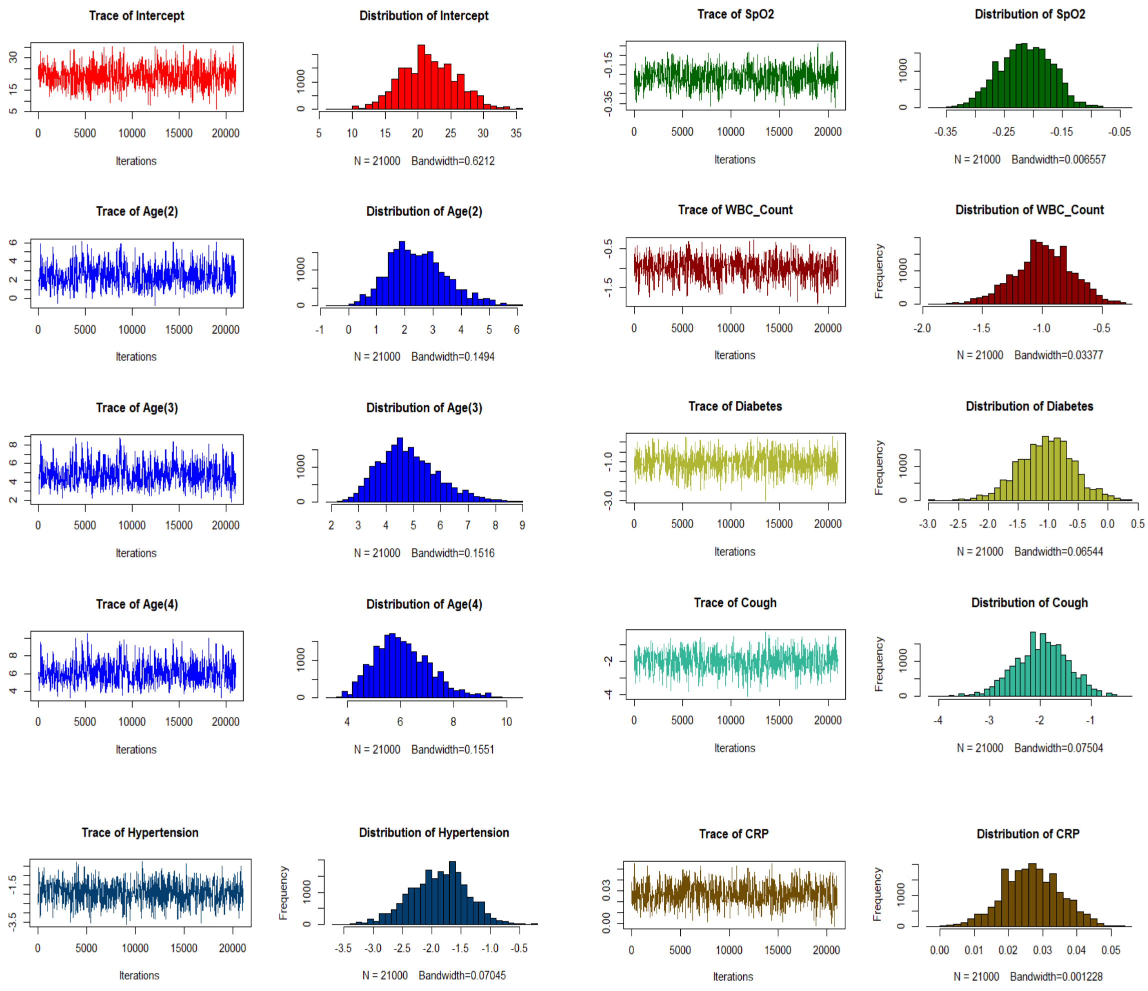
Variables	2.50%	25%	50%	75%	97.50%
(Intercept)	24.4444	32.23747	37.15356	41.72098	51.58415
Age (18-44)	0.5287	1.86368	2.52554	3.26089	4.96312
Age (45-65)	3.4288	4.62444	5.41182	6.20153	7.90676
Age (>65)	4.482	5.9219	6.74958	7.56156	9.35048
SpO2	-0.5124	-0.41129	-0.3613	-0.31581	-0.2351
WBC_Count	-1.7747	-1.43344	-1.23675	-1.04139	-0.72505
Diabetes (Yes)	2.5318	1.75568	1.37872	1.06351	0.40886

Table 11 (Continued.)

Cough (Yes)	3.5371	2.69687	2.24775	1.79789	1.05507
Hypertension (Yes)	3.4463	2.60355	2.20743	1.83704	1.11822
CRP	0.005	0.01708	0.02368	0.03057	0.04452

Figure 14:

Trace plots and Histogram charts of posterior distribution estimated for Bayesian logistic regression model (MCMC)



There are additional ways to summarize our findings such as the minimum, maximum, mean, and SD. Table 12 shows some basic statistical measure of the posterior distribution and enables us to understand the distribution easier.

Because we used a normal distribution as a prior distribution, the mean and standard deviation outputs are extremely useful in this context, particularly for our parameters b0 and b1. Keep in mind that our prior distribution had a mean of zero and a precision of 0.0001, yielding a standard deviation of 100. One major advantage of the Bayesian method over other methods such as maximum likelihood and least squares is the ability to incorporate existing knowledge into your analysis.

Table 12:

Descriptive Statistics (Minimum, Maximum, Mean, Median, SD) of the posterior findings

Variables	Mean	Meadian	SD	Minimum	Maximum
(Intercept)	27.6297	21.4559	4.2894	6.1509	35.9283
Age (18-44)	2.4636	2.3519	1.0315	-0.7125	6.1421
Age (45-65)	4.8139	4.6886	1.0792	1.8271	8.8339
Age (>65)	6.0791	5.9705	1.0708	3.2534	10.4790
SpO2	-0.2132	-0.2129	0.0453	-0.3733	-0.0390
WBC_Count	-0.9886	-0.9874	0.2441	-1.9287	-0.2663
Diabetes (Yes)	1.0398	1.0305	0.4518	0.3077	2.9687
Cough (Yes)	1.9824	1.9632	0.5305	0.2308	4.1327
Hypertension (Yes)	1.8886	1.8610	0.4864	0.2638	3.6134
CRP	0.0268	0.0267	0.0085	-0.0031	0.0547

Model Accuracy and Diagnosis Assessment

Checking the adequacy of the regression model is just as essential for logistic models as it is for general linear models. Examining the goodness-of-fit is a simple but important diagnostic tool for determining whether our model is adequate. There are two common statistical methods for determining a logistic regression model's goodness-of-fit. The first statistic is the Pearson X^2 statistic, which is calculated using observed (o) and expected, fitted, or predicted (e) observations. The other one is the G^2.

Table 13:

Test of Goodness-of-fit for the final model

Statistic	Value	df	P-value
Hosmer–Lemeshow ($\hat{C}$)	7.653	8	0.607
Deviance (G^2)	120.11	338	0.000
Nagelkerke R-Sqaure	85.21%		

Low values indicate that the model is a better fit to the data in both cases, means the observed and fitted values are similar. To evaluate the model's fit, goodness of fit statistics was computed as indicated in Table 13. The Hosmer-Lemeshow statistic was not significant, indicating that there was no evidence of model is fit, and the logistic analogue of R2 stated that about 85% of the uncertainty in the presence of no surviving from Covid19 could be of the Age, SpO2, WBC, Diabetes, Cough, Hypertension and CRP variables.

Similar to univariate models, classification accuracy was also applied on both trained (in-sample) and tested (out-sample) datasets and we can notice that in-sample prediction's area under the curve was higher than the out-sample prediction and which is corrected since the model had already seen the data. However, 83.1% overall accuracy was also considered as good model for predictions. Table 11 shows very important results in terms of models and the AUC test where provided significant and very tight range of 95% CI for both samples. In addition to that, the model also had high sensitivity and specificity for out-sample dataset with 85.568% and 79.487%, which was another sign of the model that can be reliable while predicting for unseen data points. Figure 16 explores ROC curve and shows performances of both models.

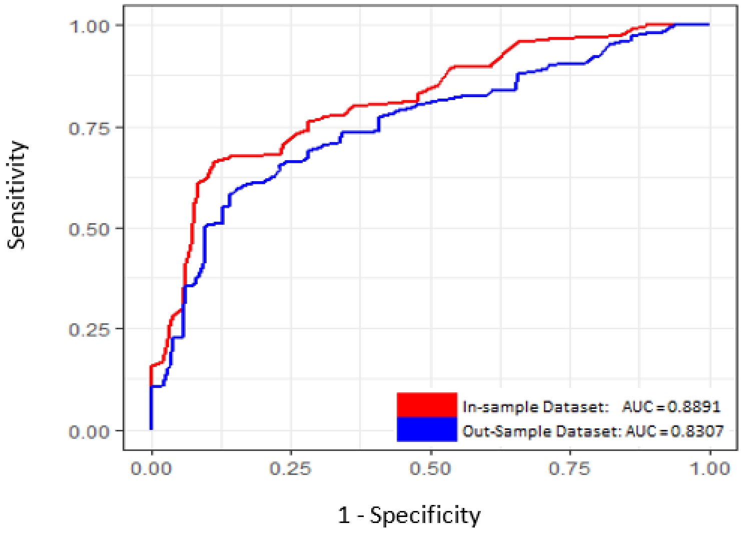
Table 14:

Confusion Matrix Result of Bayesian Logistic Regression

	Out-Sample Dataset		In-Sample Dataset	
	Recovered	Died	Recovered	Died
Recovered	95	16	103	9
Died	16	62	12	65
AUC	0.8301		0.8891	
P-Value	0.0147		0.0318	
95% CI	(0.817, 0.854)		(0.86, 0.91)	
Sensitivity	85.586%		91.964%	
Specificity	79.487%		84.416%	

Figure 15:

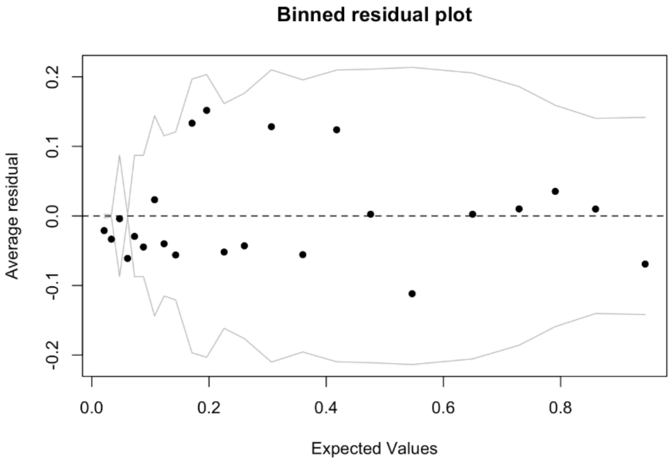
ROC Curve characteristics for Out-sample and In-sample dataset run in Bayesian Logistic Regression Model



The model appears practical as shown in Figure 16, but the residuals have some outliers, 23 binned residuals but 3 outliers = 0.13. The model performs well when fitted values are greater than about.1, but struggles below.1, where we find three negative outliers. This means that the model predicts a higher average rate of died cases in these bins than is actually the case.

Figure 16:

Bayesian Logistic Regression Assessment Graph



Implementing Other Machine Learning Methods

Before applying models such as Bayesian neural network (BNN), backpropagation neural networks (BPNN), Naïve Bayes, SVM, and kNN to the cereal dataset, it is crucial to scale the data. Data scaling is necessary because the scale of a variable alone can significantly impact the prediction variable, and using unscaled data may result in meaningless outcomes. Various techniques are commonly employed to scale data, including min-max normalization, Z-score normalization, median, and tan-h estimators.

Min-max normalization is a method that transforms the data into a common range, effectively eliminating the scaling effect caused by individual variables. Unlike Z-score normalization and median, the min-max method preserves the original distribution of the variables. In this case, we utilize min-max normalization to scale the data. Additionally, the chosen best model obtained through the stepwise forward method consists of a hidden layer with 7 neurons.

To evaluate the models and provide an overview of the error magnitudes, three key effectiveness measures are employed. One of these measures is the mean absolute error (MAE), which assesses prediction accuracy. The mean squared prediction error (MSPE) is utilized to calculate the variance between predicted and observed results. From an operational standpoint, the percentage of underestimated cases is also analyzed. In Equations (4.1 - 4.3), $\hat{y}_i$ represents the predicted values and y_i represents the observed values. These measures help gauge the performance of the models and provide insights into their accuracy.

1. Absolute Mean Square:

$$AME = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \qquad 4.1$$

2. Mean Square prediction error:

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \qquad 4.2$$

3. The proportion of underestimated prediction:

$$UP = \frac{1}{n}\sum_{i=1}^{n} U_i\, , if\ \hat{y}_i - y_i < 0, U_i = 1, 0\ otherwise \qquad 4.3$$

Table 15 reviews the five model’s testing performances and the Random Forest model did not outperform the other four fitted models, but it had somewhat smaller percentage of underestimated predictions than the SVM model. Furthermore, looking at Naïve Bayes and RF techniques, their performances were very close to BPNN. This appears to support the notion that neural network and SVM models can better approximate nonlinear functions. For testing MSPE values, the BPNN and Naïve Bayes models perform similarly. When compared to the other two models, the BNN performs the best. Bayesian methods used the hybrid Monte Carlo algorithm to update network parameters and develop neural network generalization without compromising nonlinear approximation.

Table 15:

Model Performance Analysis of Applied Algorithms

	BNN	BPNN	RF	SVM	Naïve Bayes	kNN
MAE	0.17	0.18	0.26	0.25	0.23	0.28
MSPE	0.23	0.24	0.31	0.33	0.29	0.38
Measure of Underestimation	8.10%	10.24%	11.21%	12.94%	11.74%	14.35%

Result of Bayesian Neural Network and Classic Neural Network Analysis

Since the results of the above measurements were quite similar to both BNN and BPN, thus, the confusion matrix for (BNN) and (BPNN) was calculated and presented below. Following the architecture shown in Figure 17, it was obvious that Bayesian neural network approach achieved way better results with AUC value 84.66% (95% CI of 83.26% - 85.21%) than point estimation neural network with AUC value 81.38% (95% CI of 80.25% - 82.03%).

Furthermore, both sensitivity and specificity were higher in Bayesian approach compared to classic neural network (see Table 16). Figure 18 illustrates the performance of both techniques and it can be seen that Bayesian neural network was above the classic neural network, which confirms its preferability of the model.

Table 16:

Confusion Matrix Result of Bayesian Neural Network and Classic Neural Network

	Bayesian Neural Network		Neural Network	
	Recovered	Died	Recovered	Died
Recovered	98	13	93	16
Died	16	62	19	60
AUC	0.8466		0.8138	

Table 16 (Continued.)

P-Value	0.001	0.0382
95% CI	(0.8326, 0.8521)	(0.8025, 0.8203)
Sensitivity	85.96%	83.04%
Specificity	82.67%	78.95%

Figure 17:

Bayesian Neural Network Architecture for The Study

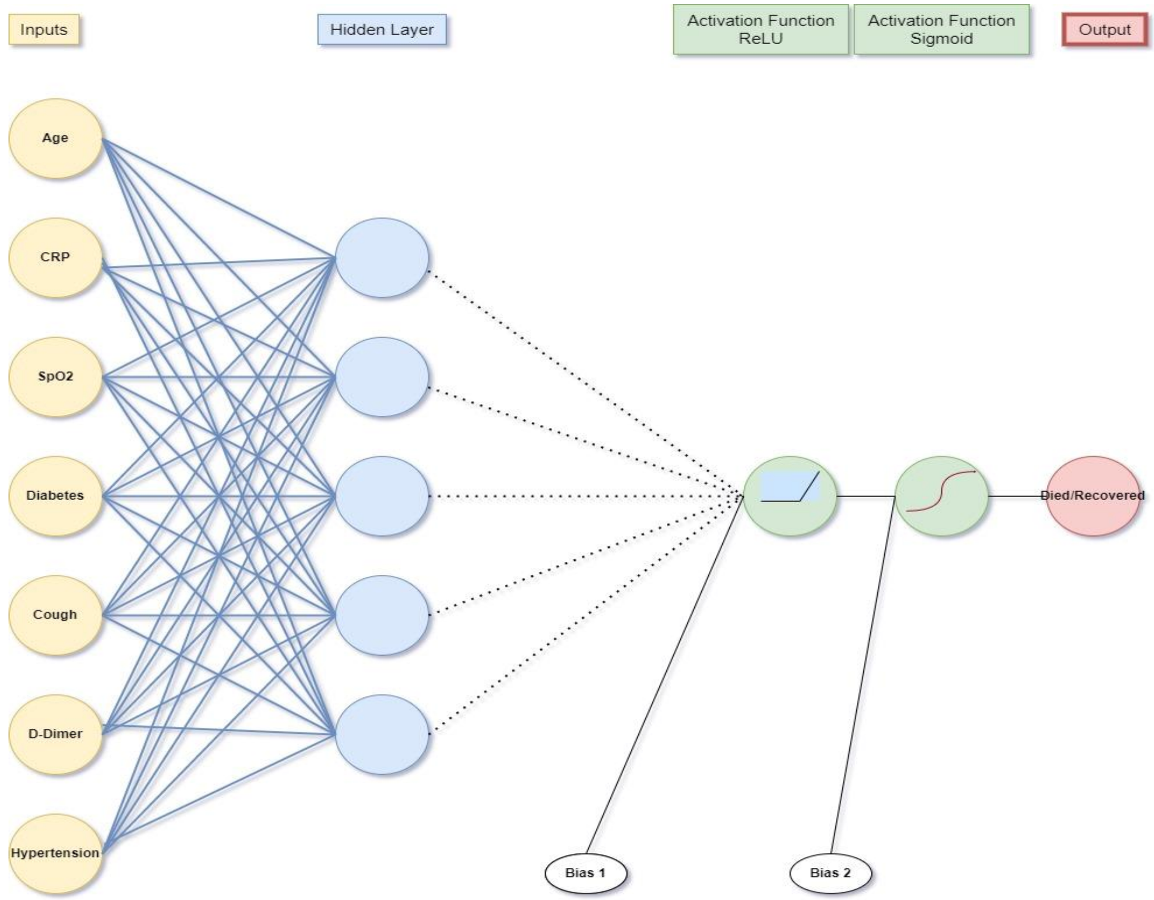
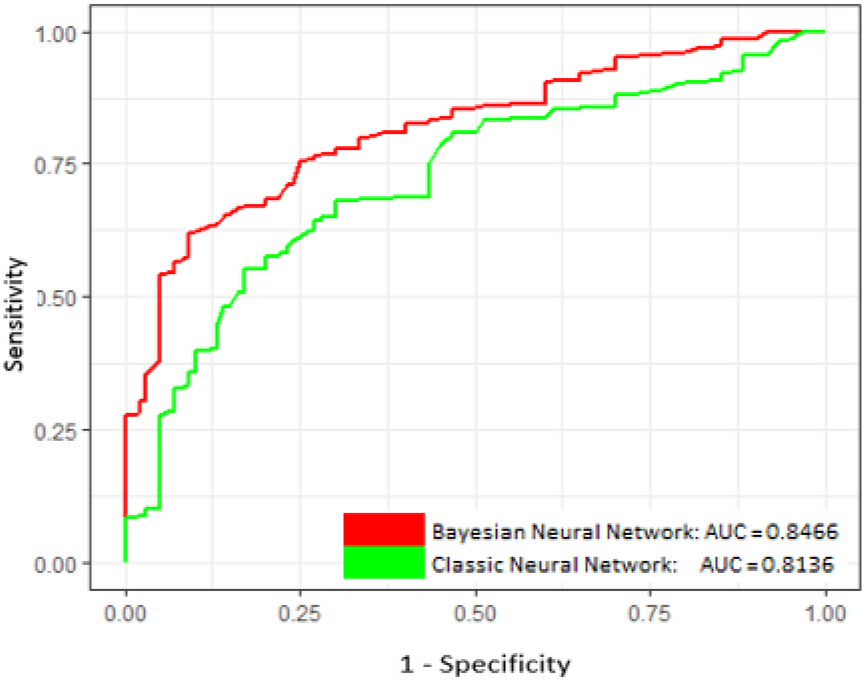


Figure 18:

ROC plot of Bayesian Neural Network and Classic Neural Network



Relative Importance Analysis of Bayesian Neural Network

The mean of the posterior distribution was calculated in Table 17 and describes the joining weight matrices of 7x7 (input-hidden) and 7x1 (hidden-output) imported from trained Bayesian neural networks with the superlative fit.

The computation of the contribution of each input variable to the output involves multiplying the input-hidden weight with the hidden-output weight. As indicated in Table 14, the magnitude and direction of the connection weights play a crucial role in determining the relative contribution of each input variable. Variables with higher connection weights signify stronger signal transfer and therefore hold greater importance in predicting death cases compared to variables with lower weights. This finding also reveals that negative values for the input variables "SpO2, WBC, CRP" correspond to lower values, which are typically associated with an increased risk, while positive values for other factors are positively associated with incident death, as outlined in Table 18.

Table 17:

Posterior Weight Matrix of Bayesian Neural Network

Input Variables	Hidden Layer1	Hidden Layer2	Hidden Layer3	Hidden Layer4	Hidden Layer5	Hidden Layer6	Hidden Layer7
Age	5.6716	-1.5890	-2.0012	-1.4991	-1.1888	1.6994	-0.7911
SpO2	1.9448	1.7333	-0.9148	7.1883	5.8809	0.3827	-0.1980
WBC_Count	0.5226	1.3398	-5.0120	-1.5861	-0.5607	-3.9801	-1.1508
Diabetes	0.4938	-15.3042	2.0612	3.4254	-0.1351	-1.7945	-0.4281
Cough	-1.5258	-5.7740	0.9962	-4.7652	-1.1682	2.3727	-1.7079
Hypertension	-1.1953	-2.1020	1.4023	0.7178	-0.3395	0.2146	1.0262
CRP	0.1698	3.9359	-0.3914	32.5188	-1.4921	-0.4147	2.1442
Multiply							
Output Layer	1.6100	-0.7695	-0.5490	-0.5840	-1.8736	-0.5930	1.5397

Input Variables	Hidden Layer1	Hidden Layer2	Hidden Layer3	Hidden Layer4	Hidden Layer5	Hidden Layer6	Hidden Layer7
Age	9.1312	1.2227	1.0986	0.8755	2.2274	-1.0077	-1.2180
SpO2	3.1311	-1.3337	0.5022	-4.1980	-11.0186	-0.2269	-0.3048
WBC_Count	0.8414	-1.0309	2.7513	0.9263	1.0506	2.3602	-1.7719
Diabetes	0.7949	11.7760	-1.1315	-2.0004	0.2532	1.0642	-0.6592
Cough	-2.4565	4.4429	-0.5469	2.7829	2.1888	-1.4070	-2.6297
Hypertension	-1.9245	1.6174	-0.7698	-0.4192	0.6361	-0.1272	1.5800
CRP	0.2734	-3.0285	0.2148	-18.9909	2.7957	0.2459	3.3014

Table 18:

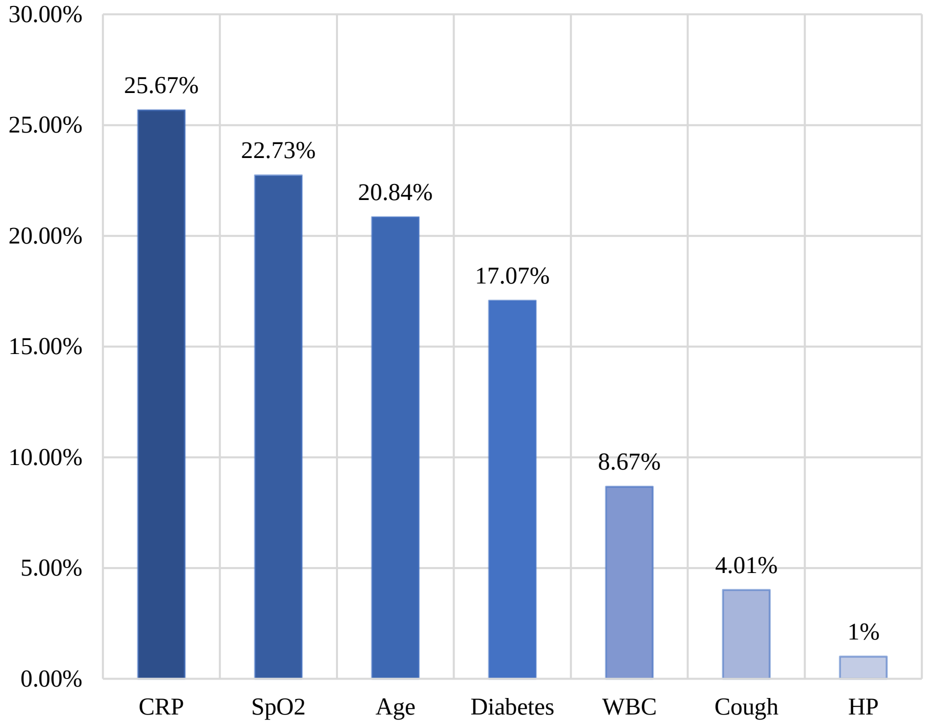
Result of Relative Importance Measurement

Input Variables	Age	SpO2	WBC	Diabetes	Cough	HP	CRP
Relative Importance	12.33	13.45	5.13	10.10	2.37	0.59	15.19
Relative Importance %	20.84%	22.73%	8.67%	17.07%	4.01%	1%	25.67%

To facilitate the interpretation of relative importance, the contributions of each input variable to the output are divided by the sum of all contributions and presented as percentages, as depicted in Figure 19. In comparison to the other factors, the CRP, SpO2, Age, and Diabetes are the strongest predictors of increasing chances to die due to Covid-19.

Figure 19:

Relative Importance of Inputs with Analyzing Weight Matrix



Confusion Matrix Analysis of Applied Models

It has come to the stage where all applied methods can be summarized into reasonable findings in order to understand and easy to follow on the model's performances. Table 19 illustrates the potential key characteristics extracted from the model after tested with unseen datasets which was the core objective of this study. To start with, AUC measurement is a performance metric for machine learning classification models that is defined as the ratio of true positives and true negatives to all positive and negative observations. Bayesian neural network had the highest value with 84.66%, followed by Bayesian logistic regression with 83.07%, classic neural network with 81.38% and Logistic regression (MLE) with 80.95%. kNN and Naïve Bayesian classification techniques were found to be the worst out of the eight methods with only success rate of overall prediction by 52.38% and 56.08% respectively. However, accuracy cannot judge the model's performance alone and there are other measurements such as F1-score, precision as well as recall.

Table 19:

Confusion Matrix Outputs of Computed Models

Classifier	AUC	F1	Precision	Recall
kNN	0.5238	0.5755	0.5701	0.581
SVM	0.6984	0.7397	0.7941	0.6923
Random Forest	0.7326	0.7788	0.8224	0.7395
Classic Neural Network	0.8138	0.8416	0.8532	0.8304
Bayesian Neural Network	0.8466	0.8711	0.8829	0.8596
Naïve Bayes	0.5608	0.5561	0.5909	0.5253
Logistic Regression (MLE)	0.8095	0.8378	0.8455	0.8304
Bayesian Logistic Regression	0.8307	0.8559	0.8559	0.8559

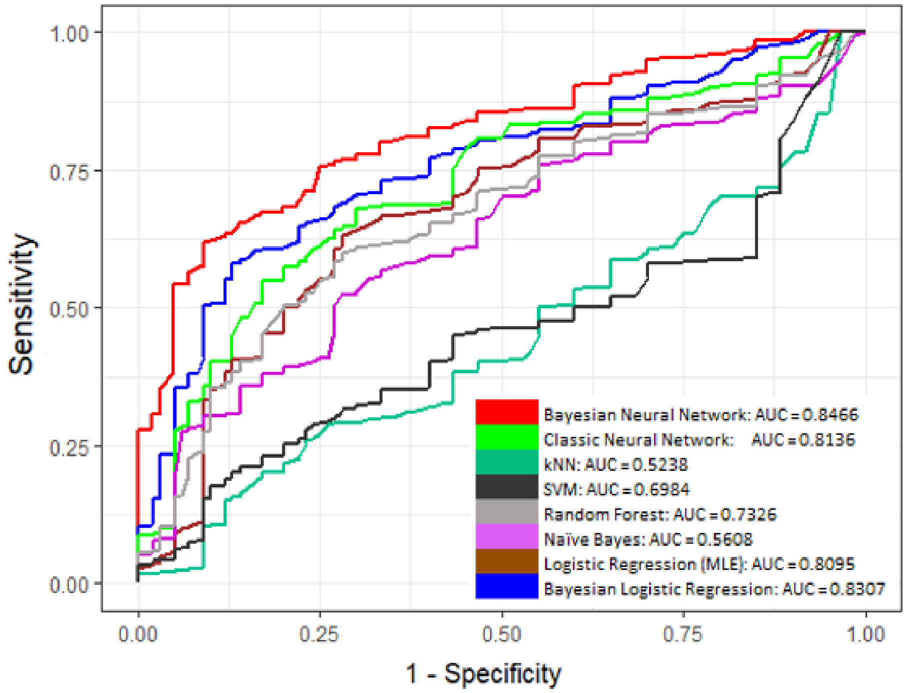
Thus, both neural network approach scored the highest precision rates with 88.29% and 86.24% for Bayesian neural network and classic neural network respectively, then followed by logistic regression (MCMC) and MLE with success rate in predicting positive records by 85.59% and 84.55% correspondingly. It was important to report that Random Forest algorithm had relatively high precision percentage with 82.24%. Moreover, in relation to recall perspective measurement, Bayesian neural network, classic neural network, logistic regression (MCMC) and logistic regression (MLE) all had similar recall rate close to 85%. This denotes the model's capability to appropriately foresee positives from definite positives. This contrasts with precision, which computes the number of positive predictions made by models out of all positive predictions made.

F1-Score was also calculated for all eight methods. It was more useful than accuracy, since an uneven class distribution was presence in our case. Bayesian neural network came out to be on the top of the list with resulting the highest F1-score by 87.11%. Logistic regression (MCMC) turned out to have the second highest F1-score rate by 85.59%. This can be interpreted as the model’s capacity to both catch positive cases and be precise with the cases.

Figure 20 displays the ROC plot of the methods where the performances can easily be detected and followed. Because ROC curves can be misleading in imbalanced datasets as in this case, precision and recall figures are frequently used instead, where the number of true positive labels is very different from the number of true negative labels.

Figure 20:

ROC Evaluation Curve of all eight models applied to predict risk of dying from Covid-19



Summary

We utilised a Bayesian approach that made use of the MCMC method in order to estimate parameters and decide whether or not the projected parameters could be recognised in a singular way. Before applying the approach to any calculations, the Bayesian method insisted that we should fulfil an initial requirement of lowering the sum of squared residuals (SSRs). For the purpose of accomplishing the work of minimization, the optimize/minimize module in Python is utilised; nevertheless, other approaches are alternatives that are just as acceptable. The goal of the MCMC technique is to get samples from the posterior PDF that are representative of the whole distribution. It is of the utmost importance to determine the correct values for the hyper parameters of both dimensions. The Markov chain Monte Carlo method is something that we used to provide parameter estimates along with confidence intervals. Utilizing MCMC affords us the opportunity to do correlation analysis between several parameters, which is an additional advantage.

CHAPTER V

Discussion And Analysis

Within the context of a Bayesian methodology, the primary concentration of this work is on the integration of Bayesian neural networks, classic neural network, Bayesian logistic regression and linear regression. It is possible to make it better by converting the predictions into a prior distribution and utilising them as prior knowledge in the Bayesian inference process by exchanging the predictions from neural networks with predicted values for the linear regression. This will result in an improved outcome. In this particular instance, we utilised a two-stage strategy; however, a combined approach, such as spatio-temporal recurrent neural networks that are able to provide accurate result predictions in the presence of uncertainty, would be the method of choice in an ideal world. Additionally, the treatment for the Covid-19 infection is only a temporary fix utilising the technique that is advised. In order to evaluate the adaptability of the model, you should use data that spans a longer period of time and a variety of spatial scales.

After adjusting for a few other variables, it was found that age had a significant correlation with the state the patients were in. As you get older, your odds of living become less likely. Age has been recognised as the primary variable in Covid-19 patients as the primary variable impacting the outcome ever since the beginning of the pandemic. The early Chinese records suggest that the case-fatality rate (CRF) rises considerably beyond the age of 60, reaching 14.8% in those over the age of 80. The patient data also revealed a significant increase in the number of patients who passed away. According to the findings of this experiment, people in age group (45-65) had 123 more odds to die as well as 436 times more chance in age group more 65 years old than patients were less than 18 years old. It can be noticed that younger people had higher chance to survive from the disease. This is in line with previous research that has established age as a significant factor for cases died because of Covid-19, particularly for individuals between the age of 45 and 64, and especially those over the age of 65 (Gralinski & Menachery, 2020; J. Wu et al., 2020). Other reports have also noted that patients in ICUs tend to be older than those who are not, and that case fatality rates are higher among older individuals (N. Chen et al., 2020; Huang et al., 2020; D. Wang et al., 2020; Yang et al., 2020). As a result, the risk of death is significantly increased in patients who are older than compared to those who are younger.

In this study, it was discovered that fever, cough, and sputum were prevalent symptoms among Covid-19 patients, particularly in those who were severely or critically ill and resulted to death. Interestingly, fever and cough are also the most common symptoms seen in patients with severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS), which are also caused by coronaviruses. Fever is known to be a primary symptom for cytokine storms, which occur when high concentrations of cytokines cause an abnormally excessive immune response and inflammation. The vital signs of severely ill patients indicated higher body temperature and respiratory rate, as well as lower SpO2 upon admission. During the outbreak, glucocorticoids have been used when SpO2 levels fall below 90%, and the oxygenation saturation index is associated with both ARDS severity and increased mortality. Our result indicated that SpO2 is associated with a 20% (1 – 0.80 = 0.20) reduction in the relative risk of dying. Thus, for a 1-unit increase in the corresponding oxygen’s level of patient admitted to hospital due to the disease is associated with a lower risk of dying due to Covid-19.

The severity of Covid-19 patients was found to be unaffected by gender in our study. While initial reports from other countries suggested a higher proportion of men experiencing severe cases of Covid-19, more recent studies have shown that similar proportions of men and women are being admitted to ICUs (N. Chen et al., 2020; Huang et al., 2020; D. Wang et al., 2020; Yang et al., 2020), indicating that any gender differences may have diminished with the increase in incidence. It's possible that earlier reports included a higher number of males due to their higher occupational risk of infection in crowded places like markets and congregations (D. Wang et al., 2020).

Our study discovered that patients who passed away due to Covid-19 had more pronounced damage to white blood cells and immune cells, and the odds ratio was estimated to be 0.37; for every 1 increase in white blood cell count, an infected Covid19 person had 63% chance to survive than not with a 95% CI of (0 and 0). Covid-19 may lead to reduced levels of T lymphocytes, including CD4+ T and CD8+ T cells, which can result in decreased production of interferon-gamma (IFN-γ), potentially contributing to disease severity (G. Chen et al., 2020). Additionally, while a more intense inflammatory response was indicated by much higher levels of inflammatory markers, such as C-reactive protein (Jousilahti et al., 2001; Sproston & Ashworth, 2018), CRP was also one of the attributors and that it was associated with a 2.72% rising the chance to die. Although the effect size small, it was statistically significant with p-value < 0.001.

D-dimer is a protein fragment that is produced in the blood after the breakdown of a blood clot through fibrinolysis (Jacobs et al., 2016). In healthy individuals, D-dimer is typically not detectable in the bloodstream, except in cases where blood clots are formed and broken down. This makes a D-dimer serum test useful for ruling out thrombotic episodes and aiding in the early diagnosis of various thromboembolic conditions, including deep vein thrombosis, pulmonary embolism, and disseminated intravascular coagulation (Le Gal & Bounameaux, 2005; Schaefer et al., 2017).

Initial studies suggested that Covid-19 patients may experience a hypercoagulable state, as evidenced by thromboembolism formation observed in pathological studies from autopsies or biopsies (Lang et al., 2003; Xu et al., 2020). Based on these findings, several researchers have linked the increase in D-dimer levels to this hypercoagulable state in Covid-19 patients (Wright et al., 2020; Zhou et al., 2020). However, other researchers suggest that elevated D-dimer levels may be associated with the inflammatory response rather than the thromboembolic condition in Covid-19 patients (Yu et al., 2020). In fact, the precise mechanisms that result in elevated D-dimer levels in Covid-19 patients remain only partially understood, and further research is needed to clarify the underlying processes.

In this investigation, we constructed a series of models that were increasingly more difficult to understand by making use of Bayesian MCMC simulation techniques. The Covid-19 model's risk factors have very good discrimination power. After taking into account the Age predictor, the AUC of the Covid-19 model became severe. According to the results of a sensitivity analysis, the performance of the model was comparable when it was used to anticipate mortality based on Covid-19 risk factors. In order to better plan health policy interventions and take the necessary actions to limit the spread of the virus as much as is practically possible, public decision-makers can benefit from using models that capture the effects of diseases on mortality of cases and can indicate whether disease has an impact on the status.

In comparison to more traditional Bayesian and likelihood-based point estimation approaches, our methodology has two significant advantages. To begin, it is compatible with models and data structures of any degree of complexity and does not call for closed-form likelihoods or ad hoc distributional limits to be placed on the form of the joint prior or posterior. Standard SIR models, which are commonly constructed using stochastic ordinary differential equations, offer a perspective of the dynamics of the epidemic that is on a coarser scale.

When utilising classical logistic regression, there are three significant sources of error that have the potential to have an effect on the results and interpretations of amortised Bayesian workflows. These sources of error are: The first possible cause of a simulation gap is a misspecification of the model, followed by corrupted data. A simulation gap occurs when a model is unable to accurately represent the dynamics of the disease that is being considered, or when data collection is distorted or polluted in ways that the model does not account for. Another scenario in which a simulation gap may occur is when the data collection is distorted or polluted. By utilising the necessary model extensions that were supported by theoretical explanations and ablation investigations, we were able to resolve these difficulties. Conventional Bayesian model verification procedures, such as low posterior probability under the prior, divergent re-simulations, or insufficient posterior predictive accuracy, are often utilised to unearth any lingering misspecifications. The fact that our method does not provide any theoretical guarantees regarding upper bounds for the residual errors, on the other hand, is an essential issue that has not yet been answered.

The Monte Carlo error is the second thing that can go wrong, and it happens when trying to estimate anything using only a few different simulations. It is also known as the error of approximation, and it is a fact that all Monte Carlo algorithms take it into account. Because we may manufacture what is effectively an unending stream of synthetic data until the continually monitored forecast accuracy is satisfactory, finding a solution to this problem in this particular setting is not very difficult to accomplish. In this regard, simulation-based inference is better suited to fully utilise the potential of deep neural networks than traditional supervised learning approaches, which rely on a limited supply of labelled real data. This is because simulation-based inference is based on the assumption that there is an infinite supply of training data.

CHAPTER VI

Conclusion And Recommendation

Conclusion

The implementation of this concept is expected to make it simpler for government agencies to keep an eye out for any contagious diseases. The findings of the model can be utilised in the formulation of public health policy, such as the administration of immunisations or the implementation of preventative measures. This research makes a contribution by utilising neural network approaches to learn complicated dependencies from the data as well as a Bayesian paradigm to associate the uncertainty in the predictions. Both of these methods are described in the following sentence. Because of this, our method has the potential to produce a model that can make accurate forecasts regarding infectious diseases and contribute to the mitigation of the negative effects those diseases have.

This research contributes by utilizing neural network approaches to learn complicated dependencies from the data as well as a Bayesian paradigm to associate the uncertainty in the predictions. Thus, our method has the potential to produce a model that can make accurate forecasts regarding infectious diseases and contribute to the mitigation of the negative effects those have diseases. To accomplish this, it is essential to keep the following information about the prior distributions in mind, as it will be utilized to estimate the parameters of the model. Even though it was assumed that these prior distributions did not provide any useful information, it is nevertheless recommended to carry out a sensitivity analysis to determine level of effectiveness.

Each parameter of the model has a normal prior distribution mean, and in addition to this, a value ranging from 102 to 106 was appended to the variance of the normal distributions. A normal prior with a variance of 106 is sufficiently non-informative and generally functions well with our dataset. This conclusion was reached as a result of the findings that the posterior distributions for the regression parameters differed only slightly from one another. This suggests that the outcomes produced by our model were reliable across a broad spectrum of prior distributions. Moreover, Bayesian neural network performed better than the other three approaches in terms of accuracy and stability as well as convergence.

We utilised state-of-the-art methodologies in Bayesian neural networks. When compared to new neural network techniques, the Bayesian neural network made use of a cutting-edge sampling strategy that improved sample quality by employing parallel computing and parallel tempering MCMC. This method was employed to improve sampling. According to the findings of our analysis, it is essential to incorporate data from a unique occurrence while formulating models. Early Covid-19 data were incorporated in the inquiry studies' dataset, which resulted in a considerable improvement in the accuracy of prediction. A high level of volatility amplifies the uncertainty introduced by models and makes predicting a very difficult task. Even while machine learning algorithms give excellent predictions, the value of those predictions is limited by volatility; hence, it is crucial that models be valid. Investors would have more faith in predictions made by Bayesian neural networks with strong uncertainty quantification achieved by Bayesian inference if these networks were used. Better forecasting performance was achieved with this strategy before the Covid-19 pandemic, which is not the case now. Even in the face of high market volatility during the early stages of the Covid-19 outbreak, the results reveal that Bayesian neural networks are able to produce trustworthy forecasts with robust uncertainty quantification. This is the case even though the results were obtained.

Recommendation

The implementation of this concept is expected to make it simpler for government agencies to keep an eye out for any contagious diseases. The findings of the model can be utilized in the formulation of public health policy, such as the administration of immunizations or the implementation of preventative measures. Another suggestion is to obtain more data with much more variables such as, x-ray, MRI, medications taken by the patients.

References

Abdulkareem, K. H., Al-Mhiqani, M. N., Dinar, A. M., Mohammed, M. A., Al-Imari, M. J., Al-Waisy, A. S., Alghawli, A. S., & Al-Qaness, M. A. (2022). MEF: multidimensional examination framework for prioritization of COVID-19 severe patients and promote precision medicine based on hybrid multi-criteria decision-making approaches. *Bioengineering*, *9*(9), 457.

Akhmerov, A. (2020). Marban E. *COVID-19 and the heart. Circ Res*, *126*, 1443-1455.

Ashish, D., Hoogenboom, G., & McClendon, R. (2004). Land-use classification of gray-scale aerial images using probabilistic neural networks. *Transactions of the ASAE, 47*(5), 1813.

Bank, W. (2020). World Bank East Asia and Pacific Economic Update, Spring 2020: Preparedness and Vulnerabilities/Global Reverberations of COVID-19. In: The World Bank.

Barton, L. M., Duval, E. J., Stroberg, E., Ghosh, S., & Mukhopadhyay, S. (2020). Covid-19 autopsies, oklahoma, usa. *American journal of clinical pathology*, *153*(6), 725-733.

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian theory* (Vol. 405). John Wiley & Sons.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Biazzo, I., Braunstein, A., Dall’Asta, L., & Mazza, F. (2022). A Bayesian generative neural network framework for epidemic inference problems. *Scientific Reports*, *12*(1), 19673.

Bikdeli, B., Madhavan, M., & Jimenez, D. (2020). the IUA, Supported by the ESC Working Group on Pulmonary Circulation and Right Ventricular Function. COVID-19 and Thrombotic or Thromboembolic Disease: Implications for Prevention, Antithrombotic Therapy, and Follow-Up: JACC State-of-the-Art Review. *J Am Coll Cardiol*, *75*, 2950-2973.

Bishop, C. M. (1993). Curvature-driven smoothing: a learning algorithm for feedforward networks. *IEEE Transactions on Neural Networks*, *4*(5), 882-884.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. International conference on machine learning, Bocco, M., Ovando, G., & Sayago, S. (2006). Development and evaluation of neural network models to estimate daily solar radiation at Córdoba, Argentina. *Pesquisa Agropecuária Brasileira*, *41*, 179-184.

Brownlee, J. (2019). A gentle introduction to logistic regression with maximum likelihood estimation. *Machine Learning Mastery*.

Bucholz, E. M., Sleeper, L. A., & Newburger, J. W. (2018). Neighborhood socioeconomic status and outcomes following the Norwood procedure: an analysis of the Pediatric Heart Network Single Ventricle Reconstruction Trial public data set. *Journal of the American Heart Association*, *7*(3), e007065.

Buntine, W. L. (1994). Operations for learning with graphical models. *Journal of artificial intelligence research*, *2*, 159-225.

Bzdok, D., Krzywinski, M., & Altman, N. (2017). Machine learning: a primer. *Nature methods*, *14*(12), 1119.

Carsana, L., Sonzogni, A., Nasr, A., Rossi, R. S., Pellegrinelli, A., Zerbi, P., Rech, R., Colombo, R., Antinori, S., & Corbellino, M. (2020). Pulmonary post-mortem findings in a series of COVID-19 cases from northern Italy: a two-centre descriptive study. *The Lancet infectious diseases*, *20*(10), 1135-1140.

Chai, Y., Ko, G. G., Bailey, L., Brooks, D., & Wei, G.-Y. (2022). CoopMC: Algorithm-architecture co-optimization for Markov chain Monte Carlo accelerators. 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA),

Chandra, R., & He, Y. (2021). Bayesian neural networks for stock price forecasting before and during COVID-19 pandemic. *Plos one*, *16*(7), e0253217.

Chen, G., Wu, D., Guo, W., Cao, Y., Huang, D., Wang, H., Wang, T., Zhang, X., Chen, H., & Yu, H. (2020). Clinical and immunological features of severe and moderate coronavirus disease 2019. *The Journal of clinical investigation*, *130*(5), 2620-2629.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., & Wei, Y. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The lancet*, *395*(10223), 507-513.

Colavita, F., Lapa, D., Carletti, F., Lalle, E., Bordi, L., Marsella, P., Nicastri, E., Bevilacqua, N., Giancola, M. L., & Corpolongo, A. (2020). SARS-CoV-2 isolation from ocular secretions of a patient with COVID-19 in Italy with prolonged viral RNA detection. *Annals of internal medicine*, *173*(3), 242-243.

Cortés-Martínez, K. V., Estrada-Esquivel, H., Martínez-Rebollar, A., Hernández-Pérez, Y., & Ortiz-Hernández, J. (2022). The State of the Art of Data Mining Algorithms for Predicting the COVID-19 Pandemic. *Axioms, 11*(5), 242.

Creel-Bulos, C., Hockstein, M., Amin, N., Melhem, S., Truong, A., & Sharifpour, M. (2020). Acute cor pulmonale in critically ill patients with Covid-19. *New England journal of medicine*, *382*(21), e70.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning,

Dinar, A. M., Raheem, E. A., Abdulkareem, K. H., Mohammed, M. A., Oleiwie, M. G., Zayr, F. H., Al-Boridi, O., Al-Mhiqani, M. N., & Al-Andoli, M. N. (2022). Towards automated multiclass severity prediction approach for COVID-19 infections based on combinations of clinical data. *Mobile Information Systems*, *2022*.

Dogan, O., Tiwari, S., Jabbar, M., & Guggari, S. (2021). A systematic review on AI/ML approaches against COVID-19 outbreak. *Complex & Intelligent Systems*, *7*, 2655-2678.

Dong, E., Ratcliff, J., Goyea, T. D., Katz, A., Lau, R., Ng, T. K., Garcia, B., Bolt, E., Prata, S., & Zhang, D. (2022). The Johns Hopkins University Center for Systems Science and Engineering COVID-19 Dashboard: data collection process, challenges faced, and lessons learned. *The Lancet infectious diseases*.

Fong, S. J., Li, G., Dey, N., Crespo, R. G., & Herrera-Viedma, E. (2020). Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. *Applied soft computing*, *93*, 106282.

Gelman, A., Gilks, W. R., & Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability*, *7*(1), 110-120.

Ghoshal, B., & Tucker, A. (2020). Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *arXiv preprint arXiv:2003.10769*.

Gralinski, L. E., & Menachery, V. D. (2020). Return of the Coronavirus: 2019-nCoV. *Viruses*, *12*(2), 135.

Greenberg, E. (2012). *Introduction to Bayesian econometrics*. Cambridge University Press.

Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., & Hui, D. S. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England journal of medicine*, *382*(18), 1708-1720.

Guo, D., Pan, S., Wang, M., & Guo, Y. (2020). Hyperbaric oxygen therapy may be effective to improve hypoxemia in patients with severe COVID-2019 pneumonia: two case reports. *Undersea Hyperb Med*, *47*(2), 181-187.

Halpin, S., O'Connor, R., & Sivan, M. (2021). Long COVID and chronic COVID syndromes. *Journal of medical virology*, *93*(3), 1242.

Hameed Abdulkareem, K., Awad Mutlag, A., Musa Dinar, A., Frnda, J., Abed Mohammed, M., Hasan Zayr, F., Lakhan, A., Kadry, S., Ali Khattak, H., & Nedoma, J. (2022). Smart healthcare system for severity prediction and critical tasks management of COVID-19 patients in IoT-fog computing environments. *Computational Intelligence and Neuroscience, 2022*.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Hernández-Lobato, J. M., & Adams, R. (2015). Probabilistic backpropagation for scalable learning of bayesian neural networks. International conference on machine learning,

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359-366.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., & Gu, X. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*, *395*(10223), 497-506.

Jacobs, B., Obi, A., & Wakefield, T. (2016). Diagnostic biomarkers in venous thromboembolic disease. *Journal of Vascular Surgery: Venous and Lymphatic Disorders*, *4*(4), 508-517.

Jain, U. (2020). Effect of COVID-19 on the Organs. *Cureus*, *12*(8).

Jalali, A., Lonsdale, H., Do, N., Peck, J., Gupta, M., Kutty, S., Ghazarian, S. R., Jacobs, J. P., Rehman, M., & Ahumada, L. M. (2020). Deep learning for improved risk prediction in surgical outcomes. *Scientific Reports*, *10*(1), 9289.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jousilahti, P., Salomaa, V., Rasi, V., Vahtera, E., & Palosuo, T. (2001). The association of c-reactive protein, serum amyloid a and fibrinogen with prevalent coronary heart disease—baseline findings of the PAIS project. *Atherosclerosis*, *156*(2), 451-456.

Khan, M. A., Khan, R., Algarni, F., Kumar, I., Choudhary, A., & Srivastava, A. (2022). Performance evaluation of regression models for COVID-19: A statistical and predictive perspective. *Ain Shams Engineering Journal*, *13*(2), 101574.

Khudhur, A. M., & Kadir, D. H. (2022). An application of logistic regression modeling to predict risk factors for bypass graft diagnosis in Erbil. *Cihan University-Erbil Scientific Journal*, *6*(1), 57-63.

Kingma, D. P., Salimans, T., & Welling, M. (2015). Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, *28*.

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, *12*(4), 307-392.

Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, *4*(1), 4.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of statistical software*, *36*, 1-13.

Lampinen, J., & Selonen, A. (1997). Using background knowledge in multilayer perceptron learning. PROCEEDINGS OF THE SCANDINAVIAN CONFERENCE ON IMAGE ANALYSIS,

Lang, Z., Zhang, L., Zhang, S., Meng, X., Li, J., Song, C., Sun, L., & Zhou, Y. (2003). Pathological study on severe acute respiratory syndrome. *Chinese medical journal*, *116*(07), 976-980.

Le Gal, G., & Bounameaux, H. (2005). D-dimer for the diagnosis of pulmonary embolism: a call for sticking to evidence. *Intensive Care Med*, *31*(1), 1-2. https://doi.org/10.1007/s00134-004-2485-0

Levashenko, V., Rabcan, J., & Zaitseva, E. (2021). Reliability evaluation of the factors that influenced COVID-19 patients' condition. *Applied Sciences*, *11*(6), 2589.

Liu, S., Luo, H., Wang, Y., Cuevas, L. E., Wang, D., Ju, S., & Yang, Y. (2020). Clinical characteristics and risk factors of patients with severe COVID-19 in Jiangsu province, China: a retrospective multicentre cohort study. *BMC infectious diseases*, *20*(1), 1-9.

Liu, Y.-M., Xie, J., Chen, M.-M., Zhang, X., Cheng, X., Li, H., Zhou, F., Qin, J.-J., Lei, F., & Chen, Z. (2021). Kidney function indicators predict adverse outcomes of COVID-19. *Med*, *2*(1), 38-48. e32.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

MacKay, D. J. (1994). Bayesian nonlinear modeling for the prediction competition. *ASHRAE transactions*, *100*(2), 1053-1062.

Madjid, M., Safavi-Naeini, P., Solomon, S. D., & Vardeny, O. (2020). Potential effects of coronaviruses on the cardiovascular system: a review. *JAMA cardiology*, *5*(7), 831-840.

Merad, M., & Martin, J. C. (2020). Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. *Nature reviews immunology*, *20*(6), 355-362.

Morin, L., Savale, L., Pham, T., Colle, R., Figueiredo, S., Harrois, A., Gasnier, M., Lecoq, A.-L., Meyrignac, O., & Noel, N. (2021). Four-month clinical status of a cohort of patients after hospitalization for COVID-19. *Jama*, *325*(15), 1525-1534.

Neal, R. M. (1998). Assessing relevance determination methods using DELVE. *Nato Asi Series F Computer And Systems Sciences*, *168*, 97-132.

Neal, R. M. (2012). *Bayesian learning for neural networks* (Vol. 118). Springer Science & Business Media.

Niraula, P., Mateu, J., & Chaudhuri, S. (2022). A Bayesian machine learning approach for spatio-temporal prediction of COVID-19 cases. *Stochastic Environmental Research and Risk Assessment*, *36*(8), 2265-2283.

Notari, A., & Torrieri, G. (2022). COVID-19 transmission risk factors. *Pathogens and Global Health*, *116*(3), 146-177.

Ntzoufras, I. (2009). Bayesian Modeling Using Winbugs. John Wiley& Sons. *Inc, Canada*.

Obaid, O. I., Mohammed, M. A., & Mostafa, S. A. (2020). Long short-term memory approach for Coronavirus disease predicti. *Journal of Information Technology Management*, *12*(Special Issue: The Importance of Human Computer Interaction: Challenges, Methods and Applications.), 11-21.

Organization, W. H. (2020a). Coronavirus disease 2019 (COVID-19): situation report, 73.

Organization, W. H. (2020b). Novel Coronavirus (2019-nCoV): situation report, 11.

Organization, W. H. (2021). COVID-19 weekly epidemiological update, 9 March 2021.

Pijls, B. G., Jolani, S., Atherley, A., Derckx, R. T., Dijkstra, J. I., Franssen, G. H., Hendriks, S., Richters, A., Venemans-Jellema, A., & Zalpuri, S. (2021). Demographic risk factors for COVID-19 infection, severity, ICU admission and death: a meta-analysis of 59 studies. *BMJ open*, *11*(1), e044640.

Press, S. J., & Press, S. J. (1989). *Bayesian statistics: principles, models, and applications* (Vol. 210). John Wiley & Sons Incorporated.

Radev, S. T., Graw, F., Chen, S., Mutters, N. T., Eichel, V. M., Bärnighausen, T., & Köthe, U. (2021). OutbreakFlow: Model-based Bayesian inference of disease outbreak dynamics with invertible neural networks and its application to the COVID-19 pandemics in Germany. *PLoS computational biology*, *17*(10), e1009472.

Ramasamy, M. N., Minassian, A. M., Ewer, K. J., Flaxman, A. L., Folegatti, P. M., Owens, D. R., Voysey, M., Aley, P. K., Angus, B., & Babbage, G. (2020). Safety and immunogenicity of ChAdOx1 nCoV-19 vaccine administered in a prime-boost regimen in young and old adults (COV002): a single-blind, randomised, controlled, phase 2/3 trial. *The lancet*, *396*(10267), 1979-1993.

Ranney, M. L., Griffeth, V., & Jha, A. K. (2020). Critical supply shortages—the need for ventilators and personal protective equipment during the Covid-19 pandemic. *New England journal of medicine*, *382*(18), e41. https://doi.org/10.1056/NEJMp2006141

Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd.

Ruth, R. D. (1983). A canonical integration technique. *IEEE Trans. Nucl. Sci.*, *30*(CERN-LEP-TH-83-14), 2669-2671.

Saeed, M., Ahsan, M., Saeed, M. H., Rahman, A. U., Mehmood, A., Mohammed, M. A., Jaber, M. M., & Damaševičius, R. (2022). An optimized decision support model for COVID-19 diagnostics based on complex fuzzy hypersoft mapping. *Mathematics, 10*(14), 2472.

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science, 2*(3), 160.

Schaefer, J. K., Jacobs, B., Wakefield, T. W., & Sood, S. L. (2017). New biomarkers and imaging approaches for the diagnosis of deep venous thrombosis. *Current opinion in hematology, 24*(3), 274-281.

Schölkopf, B., Smola, A. J., & Bach, F. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Solomon, I. H., Normandin, E., Bhattacharyya, S., Mukerji, S. S., Keller, K., Ali, A. S., Adams, G., Hornick, J. L., Padera Jr, R. F., & Sabeti, P. (2020). Neuropathological features of Covid-19. *New England journal of medicine*, *383*(10), 989-992.

Sproston, N. R., & Ashworth, J. J. (2018). Role of C-reactive protein at sites of inflammation and infection. *Frontiers in immunology, 9*, 754.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

Stouten, V., Hubin, P., Haarhuis, F., van Loenhout, J. A., Billuart, M., Brondeel, R., Braeye, T., Van Oyen, H., Wyndham-Thomas, C., & Catteau, L. (2022). Incidence and risk factors of COVID-19 vaccine breakthrough infections: a prospective cohort study in Belgium. *Viruses*, *14*(4), 802.

Taquet, M., Dercon, Q., Luciano, S., Geddes, J. R., Husain, M., & Harrison, P. J. (2021). Incidence, co-occurrence, and evolution of long-COVID features: A 6-month retrospective cohort study of 273,618 survivors of COVID-19. *PLoS medicine, 18*(9), e1003773.

Taquet, M., Geddes, J. R., Husain, M., Luciano, S., & Harrison, P. J. (2021). 6-month neurological and psychiatric outcomes in 236 379 survivors of COVID-19: a retrospective cohort study using electronic health records. *The Lancet Psychiatry*, *8*(5), 416-427.

Thye, A. Y.-K., Law, J. W.-F., Tan, L. T.-H., Pusparajah, P., Ser, H.-L., Thurairajasingam, S., Letchumanan, V., & Lee, L.-H. (2022). Psychological symptoms in COVID-19 patients: insights into pathophysiology and risk factors of long COVID-19. *Biology*, *11*(1), 61.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, *81*(393), 82-86.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P. G., & Fu, H. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, *20*(6), 669-677.

Wang, C., Wang, Z., Wang, G., Lau, J. Y.-N., Zhang, K., & Li, W. (2021). COVID-19 in early 2021: current status and looking forward. *Signal transduction and targeted therapy*, *6*(1), 114.

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., & Xiong, Y. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *Jama*, *323*(11), 1061-1069.

Wang, F., Nie, J., Wang, H., Zhao, Q., Xiong, Y., Deng, L., Song, S., Ma, Z., Mo, P., & Zhang, Y. (2020). Characteristics of peripheral lymphocyte subset alteration in COVID-19 pneumonia. *The Journal of infectious diseases*, *221*(11), 1762-1769.

Wang, K., Chen, W., Zhang, Z., Deng, Y., Lian, J.-Q., Du, P., Wei, D., Zhang, Y., Sun, X.-X., & Gong, L. (2020). CD147-spike protein is a novel route for SARS-CoV-2 infection to host cells. *Signal transduction and targeted therapy*, *5*(1), 283.

Ward, J., Harwood, R., Smith, C., Kenny, S., Clark, M., Davis, P. J., Draper, E. S., Hargreaves, D., Ladhani, S., & Linney, M. (2021). Risk factors for intensive care admission and death amongst children and young people admitted to hospital with

COVID-19 and PIMS-TS in England during the first pandemic year. *medRxiv*, 2021.2007. 2001.21259785.

Wen, Y., Vicol, P., Ba, J., Tran, D., & Grosse, R. (2018). Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*.

Wichmann, D., & Sperhake, J. (2020). Lütgehetmann Met al. *Autopsy findings and venous thromboembolism in patients with COVID-19: a prospective cohort study Ann Intern Med2020. Doi*, *10*, M20-2003.

Wilhelmsen, M., Dimakos, X. K., Husebø, T., & Fiskaaen, M. (2009). Bayesian modelling of credit risk using integrated nested laplace approximations. *NR publication*, 1-25.

Wright, F. L., Vogler, T. O., Moore, E. E., Moore, H. B., Wohlauer, M. V., Urban, S., Nydam, T. L., Moore, P. K., & McIntyre Jr, R. C. (2020). Fibrinolysis shutdown correlation with thromboembolic events in severe COVID-19 infection. *Journal of the American College of Surgeons*, *231*(2), 193-203. e191.

Wu, J., Li, W., Shi, X., Chen, Z., Jiang, B., Liu, J., Wang, D., Liu, C., Meng, Y., & Cui, L. (2020). Early antiviral treatment contributes to alleviate the severity and improve the prognosis of patients with novel coronavirus disease (COVID-19). *Journal of internal medicine*, *288*(1), 128-138.

Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The lancet*, *395*(10225), 689-697.

Xu, Z., Shi, L., Wang, Y., Zhang, J., Huang, L., Zhang, C., Liu, S., Zhao, P., Liu, H., & Zhu, L. (2020). Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The lancet respiratory medicine*, *8*(4), 420-422.

Yang, X., Yu, Y., Xu, J., Shu, H., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., & Yu, T. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *The lancet respiratory medicine*, *8*(5), 475-481.

Yu, B., Li, X., Chen, J., Ouyang, M., Zhang, H., Zhao, X., Tang, L., Luo, Q., Xu, M., & Yang, L. (2020). Evaluation of variation in D-dimer levels among COVID-19 and bacterial pneumonia: a retrospective analysis. *Journal of thrombosis and thrombolysis*, *50*, 548-557.

Zhang, Y., Xiao, M., Zhang, S., Xia, P., Cao, W., Jiang, W., Chen, H., Ding, X., Zhao, H., & Zhang, H. (2020). Coagulopathy and antiphospholipid antibodies in patients with Covid-19. *New England journal of medicine*, *382*(17), e38.

Zhang, Y., & Xu, X. (2020). Predicting As x Se 1-x glass transition onset temperature. *International Journal of Thermophysics*, *41*(11), 149.

Zhang, Y., & Xu, X. (2021a). Machine learning F-doped Bi (Pb)–Sr–Ca–Cu–O superconducting transition temperature. *Journal of Superconductivity and Novel Magnetism*, *34*, 63-73.

Zhang, Y., & Xu, X. (2021b). Predicting doped Fe-based superconductor critical temperature from structural and topological parameters using machine learning. *International Journal of Materials Research*, *112*(1), 2-9.

Zhang, Y., & Xu, X. (2021c). Predicting the material removal rate during electrical discharge diamond grinding using the Gaussian process regression: a comparison with the artificial neural network and response surface methodology. *The International Journal of Advanced Manufacturing Technology*, *113*, 1527-1533.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., & Gu, X. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The lancet*, *395*(10229), 1054-1062.

Ziemba, A. (2005). Bayesian updating of generic scoring models. *Credit Scoring and Credit Control IX*, 7-9.

Appendix A:

Full package of Bayes by Backprop Algorithm in Bayesian Neural Network Application

Algorithm (3): Bayes by Backprop

```
$STEP\ ONE$:
    $Sample\ \epsilon_i \sim N(0, I), for\ i = 1\ to\ N$
$STEP\ TWO$:
    $Set\ W_i = \mu + \log\big(1 + exp(\rho)\big) + \epsilon_i$
$STEP\ THREE$:
    $Set\ \theta = (\mu, \rho)$
$STEP\ FOUR$:
    $Let\ f(W_i, \theta) = \log q_\theta(W_i) - \log p(W_i) p(X/W_i)$
$STEP\ FIVE$:
$Calculate\ the\ gradient\ with\ respect\ to\ \mu$:
    $\frac{\partial f(W_i, \theta)}{\partial \mu} = \frac{\partial f(W_i, \theta)}{\partial W_i} + \frac{f(W_i, \theta)}{\partial \mu}$
$STEP\ SIX$.
$Calculate\ the\ gradient\ with\ respect\ to\ \rho$:
    $\frac{\partial f(W_i, \theta)}{\partial \rho} = \frac{\partial f(W_i, \theta)}{\partial W_i} \frac{\epsilon}{1 + exp\ (-\rho)} + \frac{\partial f(W_i, \theta)}{\partial \rho}$
$STEP\ SEVEN$:
$Using\ the\ estimated\ expectation\ to\ update\ the\ variational\ parameters$
    $\mu \leftarrow \mu - \alpha \frac{1}{M} \sum_{i=1}^{M} \frac{\partial f(W_i, \theta)}{\partial \mu}$
    $\rho \leftarrow \rho - \alpha \frac{1}{M} \sum_{i=1}^{M} \frac{\partial f(W_i, \theta)}{\partial \rho}$
```

Appendix B:

Stepwise Forward Selection

Null Model	Baseline Deviance		460.010	
	Residual Deviance	AIC	Change in Deviance	P-Value
M1	334.030	342.030	125.980	0.0000
M2	438.730	442.730	21.280	0.0000
M3	448.220	452.220	11.790	0.0006
M4	363.380	367.380	96.630	0.0000
M5	428.830	432.830	31.180	0.0000
M6	419.040	423.040	40.970	0.0000
M7	453.950	457.950	6.060	0.0138
M8	435.210	439.210	24.800	0.0000
M9	398.590	402.590	61.420	0.0000
M10	443.090	447.090	16.920	0.0000
M11	443.860	447.860	16.150	0.0001
M13	356.760	360.760	103.250	0.0000
M14	455.540	459.540	4.470	0.0345
M15	459.880	463.880	0.130	0.7184
M16	453.190	457.190	6.820	0.0090
Two Parameters				
M1	Baseline Deviance		334.030	
	Residual Deviance	AIC	Change in Deviance	P-Value
M1_1	320.550	330.550	13.480	0.0037
M1_2	321.250	331.250	12.780	0.0004
M1_3	254.470	264.470	79.560	0.0000
M1_4	310.830	320.830	23.200	0.0000
M1_5	312.400	322.400	21.630	0.0000
M1_6	332.300	342.300	1.730	0.1884
M1_7	309.790	319.790	24.240	0.0000
M1_8	281.920	291.920	52.110	0.0000
M1_9	321.900	331.900	12.130	0.0005
M1_10	322.560	332.560	11.470	0.0007
M1_11	240.700	250.700	93.330	0.0000
M1_12	326.180	336.180	7.850	0.0051
M1_13	330.120	340.120	3.910	0.0480
M1_14	266.510	276.510	67.520	0.0000
Three Parameters				
M1_11	Baseline Deviance		240.700	
	Residual Deviance	AIC	Change in Deviance	P-Value
M1_11_1	231.060	243.060	9.640	0.0219
M1_11_2	233.040	245.040	7.660	0.0056
M1_11_3	203.530	215.530	37.170	0.0000
M1_11_4	227.260	239.260	13.440	0.0002
M1_11_5	213.120	225.120	27.580	0.0000
M1_11_6	225.110	237.110	15.590	0.0001
M1_11_7	207.480	219.480	33.220	0.0000
M1_11_8	230.410	242.410	10.290	0.0013
M1_11_9	235.480	247.480	5.220	0.0223

M1_11_10	238.280	250.280	2.420	0.1198
M1_11_11	231.060	243.060	9.640	0.0019
M1_11_12	192.960	204.960	47.740	0.0000
Four Parameters				
M1_11_12	Baseline Deviance		192.960	
	Residual Deviance	AIC	Change in Deviance	P-Value
M1_11_12_1	188.060	202.060	4.900	0.0269
M1_11_12_2	177.070	191.070	15.890	0.0001
M1_11_12_3	181.570	195.570	11.390	0.0007
M1_11_12_4	190.650	204.650	2.310	0.1285
M1_11_12_5	167.120	181.120	25.840	0.0000
M1_11_12_6	176.790	190.790	16.170	0.0001
M1_11_12_7	168.250	182.250	24.710	0.0000
M1_11_12_8	188.000	202.000	4.960	0.0259
M1_11_12_9	190.280	204.280	2.680	0.1016
M1_11_12_9	182.780	196.780	10.180	0.0014
Five Parameters				
M1_11_12_5	Baseline Deviance		167.120	
	Residual Deviance	AIC	Change in Deviance	P-Value
M1_11_12_5_1	164.630	180.630	2.490	0.1146
M1_11_12_5_2	154.440	170.440	12.680	0.0004
M1_11_12_5_3	156.070	172.070	11.050	0.0009
M1_11_12_5_4	154.730	170.730	12.390	0.0004
M1_11_12_5_5	154.640	170.640	12.480	0.0004
M1_11_12_5_6	162.940	178.940	4.180	0.0409
M1_11_12_5_7	162.630	178.630	4.490	0.0341
Six Parameters				
M1_11_12_5_2	Baseline Deviance		154.440	
	Residual Deviance	AIC	Change in Deviance	P-Value
M1_11_12_5_2_1	147.860	165.860	6.580	0.0103
M1_11_12_5_2_2	136.840	154.840	17.600	0.0000
M1_11_12_5_2_3	143.650	161.650	10.790	0.0010
M1_11_12_5_2_4	151.900	169.900	2.540	0.1110
M1_11_12_5_2_5	150.800	168.800	3.640	0.0564
Seven Parameters				
M1_11_12_5_2_2	Baseline Deviance		136.840	
	Residual Deviance	AIC	Change in Deviance	P-Value
M1_11_12_5_2_2_1	130.05	150.05	6.790	0.0092
M1_11_12_5_2_2_2	130.12	150.12	6.720	0.0095
M1_11_12_5_2_2_3	131.58	151.58	5.260	0.0218

Appendix C:

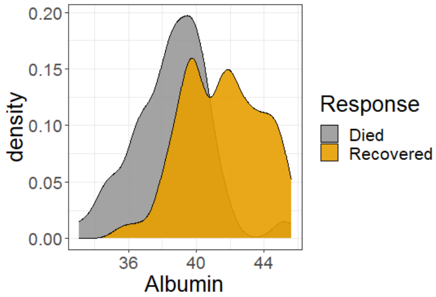
Results of Fitted Models

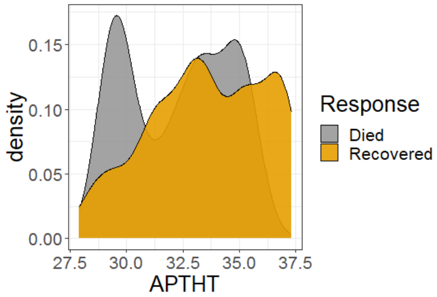
Models	Variables	Coefficients	SE	P-value	Coefficients	SE	P-value
Model 1	Intercept	-3.5410	0.7173	0.0000	-3.7640	0.7920	0.0000
	Age(2)	2.0130	0.7533	0.0075	2.2280	0.8231	0.0000
	Age(3)	3.6306	0.7578	0.0000	3.8610	0.8217	0.0000
	Age(4)	5.1066	0.7751	0.0000	5.3730	0.8542	0.0000
Model 2	Intercept	29.7000	4.2250	0.0000	30.3925	4.4679	0.0000
	Age(2)	2.3096	0.8196	0.0048	2.5754	0.9084	0.0000
	Age(3)	4.5109	0.8542	0.0000	4.8306	0.9312	0.0000
	Age(4)	5.7969	0.8824	0.0000	6.1479	0.9695	0.0000
	SpO2	-0.3585	0.0462	0.0000	-0.3691	0.0489	0.0000
Model 3	Intercept	34.4009	4.8953	0.0000	35.6934	4.9764	0.0000
	Age(2)	2.4293	0.8766	0.0056	2.6596	0.9829	0.0000
	Age(3)	4.4973	0.9250	0.0000	4.7642	1.0152	0.0000
	Age(4)	6.0688	0.9996	0.0000	6.4526	1.0960	0.0000
	SpO2	-0.3551	0.0513	0.0000	-0.3692	0.0518	0.0000
	WBC_Count	-1.1794	0.2036	0.0000	-1.2367	0.2064	0.0000
Model 4	Intercept	35.6327	5.3110	0.0000	30.5759	4.3413	0.0000
	Age(2)	2.5194	0.8929	0.0048	2.6760	0.9218	0.0000
	Age(3)	4.5427	0.9418	0.0000	4.5474	0.9539	0.0000
	Age(4)	6.0280	1.0192	0.0000	5.9553	0.9850	0.0000
	SpO2	-0.3726	0.0561	0.0000	-0.3062	0.0448	0.0000
	WBC_Count	-1.2493	0.2121	0.0000	-1.1932	0.1975	0.0000
	Diabetes(Yes)	1.5790	0.4357	0.0003	-1.4828	0.4017	0.0000
Model 5	Intercept	31.8854	5.3152	0.0000	29.3049	3.6401	0.0000
	Age(2)	1.9192	0.9389	0.0409	2.1960	1.0087	0.0000
	Age(3)	4.1428	0.9711	0.0000	4.2187	0.9836	0.0000
	Age(4)	5.3996	1.0132	0.0000	5.4661	1.0034	0.0000
	SpO2	-0.3318	0.0563	0.0000	-0.2241	0.0385	0.0000
	WBC_Count	-1.3481	0.2246	0.0000	-1.2937	0.2052	0.0000
	Diabetes(Yes)	1.4173	0.4604	0.0021	-1.2350	0.3945	0.0000
	Cough(yes)	1.6189	0.4763	0.0007	-1.6963	0.4392	0.0000
Model 6	Intercept	32.5760	5.9760	0.0000	30.4264	4.0674	0.0000

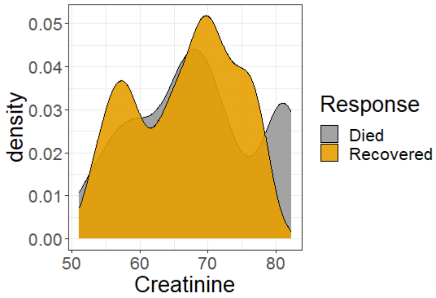
Models	Variables	Coefficients	SE	P-value	Coefficients	SE	P-value
	Age(2)	2.1677	0.9211	0.0186	2.3715	0.9638	0.0000
	Age(3)	4.5358	0.9735	0.0000	4.4284	0.9724	0.0000
	Age(4)	5.8882	1.0287	0.0000	5.8093	1.0042	0.0000
	SpO2	-0.3503	0.0640	0.0000	-0.2149	0.0425	0.0000
	WBC_Count	-1.4582	0.2385	0.0000	-1.3594	0.2110	0.0000
	Diabetes(Yes)	1.1452	0.4897	0.0194	-0.8720	0.4289	0.0000
	Cough(yes)	2.2120	0.5543	0.0001	-2.1770	0.5090	0.0000
	Hypertension(Yes)	1.9815	0.5340	0.0002	-1.7779	0.4722	0.0000
Model 7	Intercept	28.6433	6.1532	0.0000	27.6297	4.2894	0.0000
	Age(2)	2.2171	0.9668	0.0218	2.4636	1.0315	0.0000
	Age(3)	4.8459	1.0393	0.0000	4.8139	1.0792	0.0000
	Age(4)	6.0925	1.0975	0.0000	6.0791	1.0708	0.0000
	SpO2	-0.3333	0.0655	0.0000	-0.2132	0.0453	0.0000
	WBC_Count	-1.1157	0.2649	0.0000	-0.9886	0.2441	0.0000
	Diabetes(Yes)	1.2976	0.5109	0.0111	1.0398	0.4518	0.0000
	Cough(yes)	2.0109	0.5858	0.0006	1.9824	0.5305	0.0000
	Hypertension(Yes)	2.0048	0.5565	0.0003	1.8886	0.4864	0.0000
	CRP	0.0225	0.0093	0.0156	0.0268	0.0085	0.0000

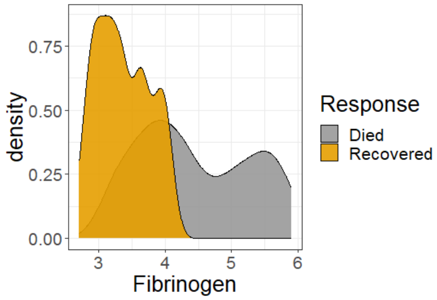
Appendix D:

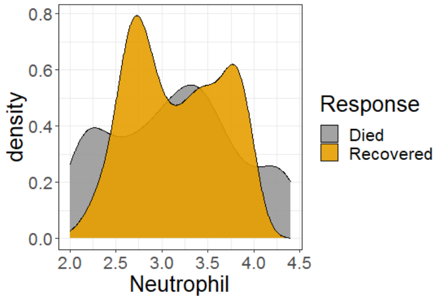
Density illustration of the non- significant variables

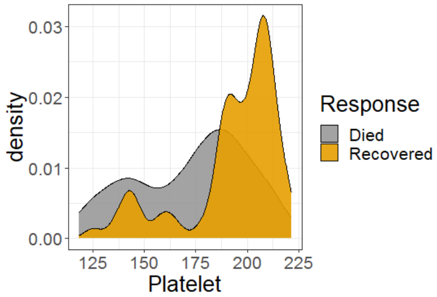


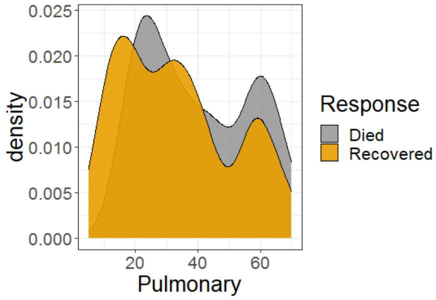


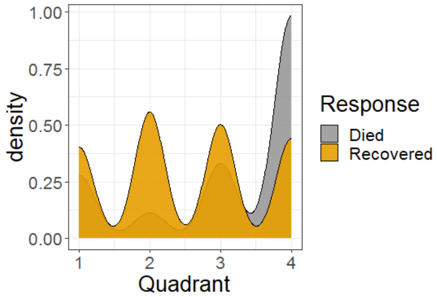


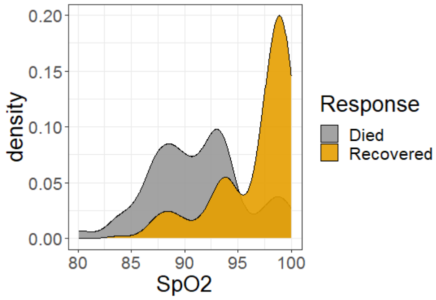






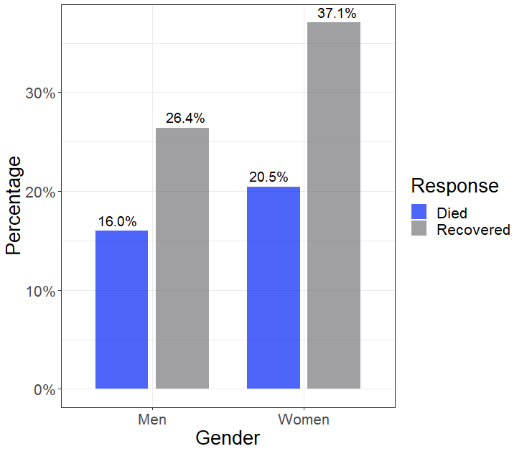


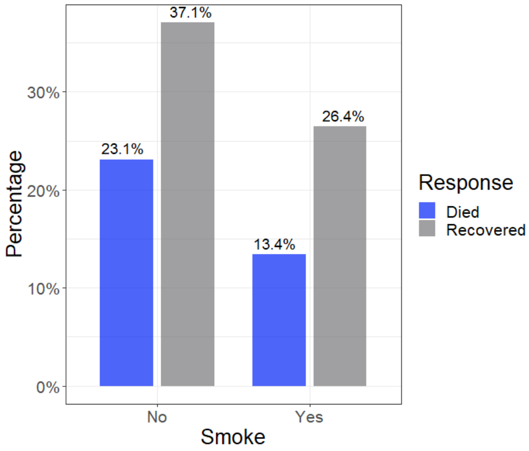


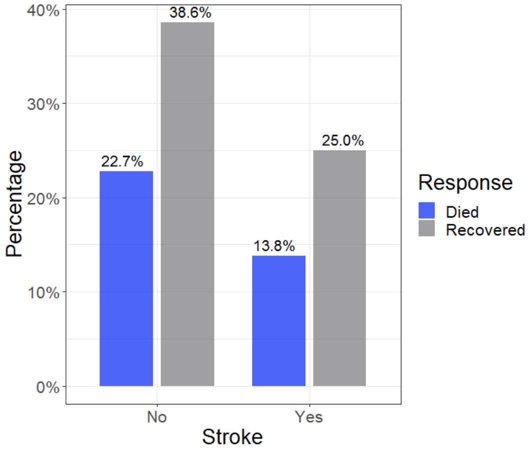


Appendix E:

Bar chart of non-significant categorical variables







Appendix F

Curriculum Vitae

Personal Information

Surname, Name:	Khidir, Hewir
Date of Birth:	30 March 1987
Place of Birth:	Erbil, Iraq

Education

Degree	Department/Program	University	Year of Graduation
M.Sc.	Statistics	University of Nottingham	2013
B.Sc.	Statistics	Salahaddin University	2009

Master Thesis Title: High Dimensional Sparse Regression.

Work Experience

Title	Place	Year
Assistant Lecturer	NEU, Faculty of Engineering, Department of Biostatistics	2014-Present
Research Assistant	Salahaddin University, College of Administrations and Economics, Department of Statistics	2009-2011

Foreign Languages

- Fluent Spoken and Written English
- Intermediate Level in Arabic

Publications In International Refereed Journals

1. Motivations Of Using Facebook As Social Networking Sites Among Students At Soran University, International Journal Of Social Science & Economic Research, Issn: 2455-8834, Volume:03, Issue:05 May 2018.

2. Healthy Life Expectancy at Birth: A Comparison Study between Developed, Developing and Undeveloped Countries, Journal of Nutrition and Internal Medicine, Vol. 23, N. 1: e2021035, 2021.
3. Logistic regression analysis of finding associated factors to predict loss weight adults in Erbil City (2018), Journal of Nutrition and Internal Medicine, Vol. 24, N. 4: e2022117, 2022.
4. Bayesian Machine Learning Analysis with Markov Chain Monte Carlo Techniques for Assessing Characteristics and Risk Factors of Covid-19 in Erbil City-Iraq 2020-2021, Alexandria Engineering Journal, September 2023, Volume78 (IssueComplete) Pages, p.162To – 174.