

https://dspace.org/

The use of machine learning techniques in identifying gender differentials in COVID-19 hospitalizations, probabilities of hospitalization outcomes and hidden correlations with demographic and clinical factors

Malaatjie, Meghan Abigail

2024
https://hdl.handle.net/10539/37622

Downloaded from a DSpace repository. https://dspace.org/

The use of machine learning techniques in identifying gender differentials in COVID-19 hospitalizations, probabilities of hospitalization outcomes and hidden correlations with demographic and clinical factors



A research report submitted to the Faculty of Health Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science in Epidemiology.

By

Meghan Abigail Malaatjie

Student number: 1155443

18 May 2023

Supervisor(s): Prof. Mary Kawonga and Prof. Bruce Mellado

School of Public Health, Faculty of Health Sciences

MSc Submission Format

This MSc research report is submitted in the format of a submissible research article. The article will be submitted to the Health Science Reports Journal. This is to certify that the article follows the submission guidelines for the Journal. The author guidelines are attached (Appendix II). The research protocol for this MSc research report was approved by the Health Sciences Postgraduate Research Committee on the 2nd November 2021 and is attached (Appendix III).

Declaration

I hereby declare that this is work I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.

Signature: __  Date: 18 May 2023

DECLARATION: Student's contribution to article and agreement of co-authors

We, the co-authors acknowledge that Meghan Malaatjie with student number 1155443 is the first author of this paper. She is responsible for the content of the paper and has done the work on the following:

- ✓ Protocol writing
- ✓ Data analysis
- ✓ Manuscript write-up

Agreement by co-authors:

By signing this declaration, we the co-authors agree to the use of this article by the student as part of her research report.

Co-author

Full name: **Thuso Mathaha**

Signature:  Date: 05/12/2022

Co-author

Full name: **Benjamin Lieberman**

Signature:  Date: 05/12/2022

Co-author

Full name: **Nidhi Tripathi**

Signature:  Date: 05/12/2022

Co-author

Full name: **Jude Kong**

Signature:  Date: 03/12/2022

Co-author

Full name: **Marie-Katherine Waller**

Signature:  Date: 02/12/2022

Co-author (MSc Co-supervisor)

Full name: **Mary Kawonga**

Signature:  Date: 02/12/22

Co-author (MSc Co-supervisor)

Full name: **Bruce Mellado**

Signature:  Date: 02/11/22

Dedication

This research report is dedicated to my Apostle, Edgar Hill Roscoe, who taught me that on the handlebars of success are blood, sweat, tears and pain. Thank you for your constant teachings on hard work, always striving to do better and putting God first in all I do.

Acknowledgement

The completion of this research came about through various individuals taking their time to support me. To each and every one of you *I thank you* for the role you played on this journey.

- To my Lord and my God, thank you for granting me the wisdom and strength to do all things.
- A heartfelt thank you, to both my research supervisors, Dr Mary Kawonga and Prof Bruce Mellado, for their patience, guidance, and insightful contributions.
- To my family, for the constant prayers, encouragement, and support. Thank you for ensuring that I was well nourished and caffeinated throughout this degree. Particularly my parents who supported me even when they did not always understand what I was doing.
- To the Africa-Canada AI & Data Innovation Consortium, for funding my studies.
- To Mr Thuso Mathaha, for his irrefutable assistance during the coding sessions in Python.
- To Mr Solly Cave, for always providing words of wisdom and encouragement when most needed.

The use of machine learning techniques in identifying gender differentials in COVID-19 hospitalizations, probabilities of hospitalization outcomes and hidden correlations with demographic and clinical factors

Meghan Malaatjie (Corresponding author)

School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa| Email: 1155443@students.wits.ac.za

Thuso Mathaha

School of Physics, Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, South Africa |Email: 1144845@students.wits.ac.za

Benjamin Lieberman

School of Physics, Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, South Africa |Email: 716034@students.wits.ac.za

Nidhi Tripathi,

School of Physics, Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, South Africa |Email: 2515639@students.wits.ac.za

Jude Kong

Department of Mathematics and Statistics, York University, Toronto, Canada | Email: jdkong@yorku.ca

Marie-Katherine Waller

Gender at Work, Toronto, Canada | Email: mariekatw@gmail.com

Mary Kawonga*

School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa | Email: mary.kawonga@wits.ac.za

Bruce Mellado*

iThemba LABS, National Research Foundation, Johannesburg, South Africa

School of Physics, Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, South Africa | Email: bruce.Mellado@wits.ac.za

***Shared senior authorship**

***Keywords**: COVID-19, Artificial Intelligence, Sex-disaggregated, Machine Learning*

Abstract

Background:

Sex-differentiated data on hospitalisation frequency, case severity, pre-existing medical conditions, and mortality outcomes amongst Covid-19 hospitalised patients is needed but limited in Gauteng province, the epicentre of the Covid-19 pandemic in South Africa. This study aims to investigate whether Machine Learning techniques can provide insight into gender differentials in COVID-19 hospitalizations throughout the four waves of the pandemic, in the Gauteng province of South Africa.

Method:

A weak supervision learning algorithm was used to perform binary classification. The training of a DNN was performed on 14 features of patient characteristics (Demographic variables, presence of co-morbidity, care received upon admission and setting of care), to separate the two classes of data sets: a) severe disease class (a proxy measure of higher severity, which included those who died during admission or were admitted into an intensive care (ICU) or high care unit (HCU)), and b) less severe disease class.

Results:

The number of Covid-19 hospitalisations was highest in wave 3 for both males and females, and higher in females than males across all 4 waves. The observed difference in COVID-19 hospitalization frequency between men and women was the highest in the 20 - 40-year age group with a ratio of 1:3. There was a higher frequency of COVID-19 hospitalization for hypertension, diabetes, and HIV frequencies across all age groups.

Conclusion:

This study demonstrated the utility of machine learning for analysing multidimensional sex-disaggregated data to provide accurate, real-time information for public health monitoring of sex-differences in the Gauteng province.

Contents

..... 1

MSc Submission Format 2

Declaration 2

DECLARATION: Student's contribution to article and agreement of co-authors 2

Dedication 4

Acknowledgement 5

Abstract 7

Introduction 10

Leveraging machine learning for Covid-19 analysis 12

Data source 14

Method 15

Pre-processing 16

Feature selection 16

Data splits 17

Evaluation metrics 17

Results 17

Machine Learning Algorithm Output 17

Hospitalization frequency by gender across 4 waves 18

Gender differences in hospitalization frequency by age 18

Gender differences in Covid-19 hospitalisation by 20

Hospitalisation frequency by geographical location 21

Discussion 22

Conclusion 25

References 26

Appendices 29

Appendix I: Evaluation Metric – Roc Curve 29

Appendix II: Ethics Approval 30

Appendix III: Journal for Health Science Reports Author Guidelines 31

Appendix IV: Approved Research Protocol 42

Introduction 44

Background 44

Literature review 45

Problem statement: 47

Justification of the research 48

Aim & objectives 48

Aim: 48
Objectives are to: 48
Methods 49
Research design and scope 49
Data source 49
Variables 50
Overview of Proposed Solutions 51
Pre-processing 51
Feature selection 51
Data splits 52
Data analysis 52
Performance testing and optimization 53
Performance evaluation 53
Environmental setup: 54
Research limitations 54
Ethical considerations 55
Project timeline 55
References 55

Introduction

The outbreak of the Corona Virus Disease 2019 (COVID-19) impacted over one hundred countries [1]. The United Nations in 2020 predicted that COVID-19 and the resultant lock downs would have disproportionately negative impacts on women considering their lower social and economic status in many societies. This has been observed during the COVID-19 pandemic globally, as women have been inordinately affected by higher Covid-19 caseloads in comparison to their male counterparts [2]. Though less attention has been drawn to the sex distribution of Covid-19 fatalities, these often show disparities geo-spatially and along socioeconomic gradients, and mortality estimates by sex vary across different population groups and contexts [3]. Differential health outcomes between males and females may be due to societal-defined differences (gender), or biological (sex) differences between males and females, the latter including differences in pharmacology and immunity [4].

Gender has been defined as the socially constructed norms, and relationships, which a given society assigns to define and structure how women, men, girls, boys, and gender-diverse people ought to behave and influences and shapes behaviours and all spheres of life [5]. Sex-differentiated data on Covid-19 and related outcomes are a useful starting point towards understanding the role of gender norms in Covid-19 risks and outcomes. In Gauteng province, the epicentre of Covid-19 in South Africa [6], multi-dimensional Covid-19 surveillance data sets are available, but innovative analytics that can provide the much-needed sex-differentiated data on Covid-19 hospitalisation and related outcomes have not been leveraged. This paper explores this gap.

The swift spread of the COVID-19 outbreak resulted in numerous countries implementing stringent public health and social measures, including lock downs. The existing gender vulnerabilities, aggravated by the breakdown in social infrastructure during lock downs, contributed to differences in exposure to, prevalence and outcomes of COVID-19 between men and women [7]. Early data has indicated that among the paramount factors which form part of the determinants that influence the susceptibility and vulnerability to COVID-19, gender and sex not only intersect with each other but also with other social determinants [8].

Gender roles and norms play a significant role in health inequities by causing differentials in vulnerability and risk, health seeking behaviour and access to healthcare, which impact on differences in health outcomes between men and women [9]. Studies have demonstrated that gender norms and roles contribute to women disproportionately taking up the burden of family care-giving, and providing care as front-line workers in healthcare settings, thus increasing their exposure risk [10]. Furthermore, at a time when many countries imposed strict lock-downs, women faced increased care-giving responsibilities, loss of work and increased gender-based violence [7]. As observed in previous

disease outbreaks, such as the Zika virus and Ebola outbreak in South America and West Africa, gendered norms and role expectations influence vulnerability, risk, and socioeconomic impacts of outbreaks in ways that disproportionately affect women [10]. Women have been identified in other health conditions as more vulnerable to poor health outcomes than men, as a result of gender-based inequities in access to income and other resources [5] and have also been shown to face reduced ability to recover from economic shocks and access protection at a social level [2].

In March 2020, a strict lockdown was imposed by the South African government with the aim of flattening the pandemic curve, thus allowing the healthcare system to make necessary preparations for the expected rise in hospital admissions amongst those infected with SARS-CoV-2, the virus that causes Covid-19 [1]. As a result of these control measures, South Africa experienced multiple layers of health and social system issues adversely affecting women's health and livelihoods [11]. The impacts on Covid-19 hospitalisation and mortality have been observed. Data from the South African National Department of Health revealed that between April 2020 and July 2021 there were observed differences between women and men in the number of Covid-19 hospitalizations - women had approximately 6% more cases of COVID-19 hospitalization than men, and 53% of the total COVID-19 hospitalizations were women [12].

A recent study by Jassat et al., on COVID-19 hospitalization in South Africa explored the juncture of age, gender, race, and socio-economic status and found that the risk of severe case of disease for COVID-19 disease was unequal among different sub-population groups and communities [13]. For example, being older, male or from a minority group, key population (i.e., homosexual men and or sex workers), or lower socio-economic background increased hospitalised Covid-19 patients' risk of having more severe COVID-19 or even dying from the disease [14]. Additionally, higher rates of COVID-19 mortality amongst those hospitalised were observed among people of colour who did not belong to the white racial group. The study provided insights showing that COVID-19 mortality was influenced by numerous intersecting risk factors including sex (male, female) and others (e.g., race, province, wave period and presence of a co-morbidity). The results from that same study also showed though the risk of Covid-19 mortality increases with age within the South African population, during March 2020 to January 2022 the COVID-19 in-hospital mortality reported among younger age groups (younger than 30 years of age) were proportionally more than that observed in higher income countries. Other Available South African data highlights pre-existing medical conditions (e.g., non-communicable diseases such as diabetes, hypertension, and cancer) as risk factors for Covid-19 and poorer outcomes [13].

Statistics South Africa's online mapping tool – the South African COVID-19 Vulnerability Index (VIndex) which came about as a means of making use of smaller area population data available while incorporating dimensions and indicators to statistically reflect local risk factors that may aid the advancement of COVID-19 infections for individuals who meet a set criteria measured – showed that approximately 33% of respondents reported that a household member had a pre-existing non-communicable disease [15]. These conditions have been shown to greatly increase the probability of serious illness and mortality in individuals who are infected with COVID-19 [16]. Recent studies elsewhere also convey that an increased risk of NCDs is a risk factor for COVID-19 hospital admission [6].

Sex-differentiated data on hospitalisation frequency, case severity, pre-existing medical conditions, and mortality outcomes amongst Covid-19 hospitalised patients is needed in Gauteng province, the epicentre of the Covid-19 pandemic in South Africa. This kind of analysis is essential to inform response strategies for the prevention and mitigation of severe health outcomes and the identification of those who are more at risk. Analysing differentials in Covid-19 hospitalisation risks and outcomes between men and women is a useful starting point towards a better understanding of gender-based inequities in Covid-19. Sex-differentials in COVID-19 can be studied by numerous analytical methods and have commonly been studied using traditional epidemiological and bio-statistical approaches. However artificial intelligence (AI) techniques have not been leveraged. AI has the potential to be a more efficient approach, if used correctly [17]. Machine learning forms part of advanced AI concepts which provide more complex and objective algorithmic techniques for multidimensional data analysis in comparison to epidemiological methods and assists with the diagnosis, detection, and monitoring of numerous diseases [18].

Leveraging machine learning for Covid-19 analysis

Machine learning is an innovative analytical technique that would be useful, given its ability to provide more precise multidimensional data analyses, the outputs of which could be used to better understand gender and other inter-sectioning factors. Machine learning therefore has much potential to fill gaps in existing data and provide information to guide more tailored COVID-19 responses that can limit the impact of the pandemic on the most vulnerable and worst affected population groups. The use of AI techniques (such as machine learning) for the development of models for predicting COVID-19 trends has aided in early detection of infection and predicting pandemic peaks which have proven to reduce the burden on healthcare systems [19]. These machine learning methods can be categorised in four different approaches namely, supervised learning, unsupervised learning, weak supervision (semi-supervised) and reinforcement learning [18]. This paper uses weak supervision

machine learning to identify gender differentials. The use of machine learning methodology in studying gender differentials in COVID-19 allows for the optimization of risk stratification as well as classifying and predicting the overall outcomes for COVID-19 patients. Such methods have great potential for learning from data and were applied in our study for understanding the gender differentials in COVID-19 hospitalizations and outcomes, with minimal pre-existing knowledge. The terms sex and gender are often used interchange-ably in the literature [5]. For the purpose of this analysis, the term gender-differentials will be used to mean sex-differentials.

That the COVID-19 pandemic quickly became a health emergency globally rapidly increased the need for collaboration across different sectors and countries. This urgent need was met with international collaboration of diverse health experts and AI-based methodologies that informed the monitoring, tracking and response to disease spread, containment and identified ways to mitigate the effect of the pandemic [17]. Moreover, the introduction of AI to public health management has displayed promising opportunities for more accurate and efficient data analysis, necessary for quick decision making in instances of disease outbreaks, as machine learning techniques learn from past data and perform real time analysis without human intervention in comparison to traditional methods of analysis. For instance, the work of Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), in partnership with the Department of health of the Gauteng Province during the COVID-19 pandemic in South Africa provided critical insights into inequalities and gender differences faced by subpopulation groups and communities in Gauteng [20]. As seen in a pilot study conducted by the ACADIC on Gauteng data, the consortium provided multi-dimensional data analysis on COVID-19 hospitalization and modelled fourteen features of individual-specific characteristics [17]. Methods of weak supervision machine learning were used to develop a classification procedure. The study utilized a deep neural network (DNN) to train separate classes within the data-set, these classes focused on the severity of the illness and grouped data into severe illness and less severe illness, respectively. The pilot study discovered that the severity of disease was enhanced by several factors which included gender, age, and co-morbidity [17]. This present study aims to build upon that work, and leverage AI to bridge the gap in in-depth understanding of gender differentials in COVID-19 hospitalizations throughout the four waves of the pandemic, in the Gauteng province of South Africa. In this paper we leverage AI to identify gender differences in hospitalization frequencies, severity and outcomes across waves 1 to 4 in Gauteng, perform a descriptive analysis of sex-dis-aggregated COVID-19 hospitalisation data and identify hidden related clinical (pre-existing conditions) and demographic variables [21].

Data source

The multidimensional data used for this analysis was obtained from the Gauteng Department of Health in South Africa. These data were collected by the National Institute for Communicable Diseases (NICD) on behalf of the South African National Department of Health, using DATCOV, a national Covid-19 hospital surveillance system [22]. The purpose of the DATCOV surveillance system was to monitor and describe trends in the epidemiology of COVID-19 hospitalizations within South Africa, including the features of patients hospitalized [8]. The surveillance data were submitted to the NICD by hospitals in all nine South African provinces from both private and public health facilities. Hospitals were requested to report for all COVID-19 admissions, patient demographic data, clinical information (inclusive of any co-morbidity patients had), clinical care received (e.g., whether received oxygen, ventilation), type of ward admitted to (intensive care, high care or general care ward), and outcomes (e.g. dead, discharged alive) [22].

The focus of our analysis in this paper is on COVID-19 hospitalizations that occurred in public and private hospitals specifically within the Gauteng province between 7 March 2020 and 25 March 2022. The focus is on Gauteng due to the province currently being the most populated in comparison to the other eight provinces within the country and experiencing the highest Covid-19 case hospitalisation load in the country [6]. The anonymized dataset that we analysed included hospitalisations across all five districts in Gauteng (three metropolitan municipalities and two district municipalities) [6]. All twenty-five health sub-districts distributed across these five districts were included in the analysis, including both impoverished and affluent suburbs. The five districts (municipalities) include: City of Tshwane Metropolitan Municipality (six sub-districts); City of Johannesburg Metropolitan Municipality (JHB – seven sub-districts); City of Ekurhuleni Metropolitan Municipality (EMM – six sub-districts); West Rand District Municipality (three sub-districts); and Sedibeng District Municipality (three sub-districts).

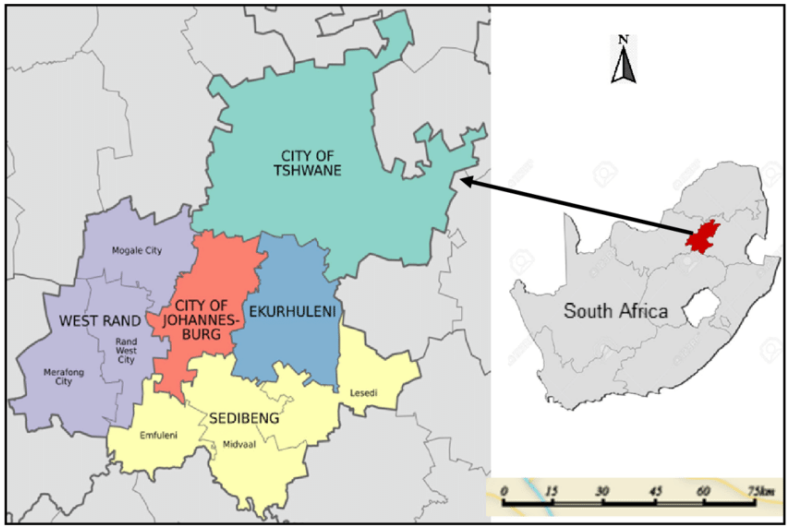

Figure 1 District map of the Gauteng province

Method

A weak supervision learning algorithm was used to perform binary classification, which created two categories within the dataset – "severe Covid-19" class which referred to patients with severe symptoms from the virus and "less severe Covid-19" class which referred to patients with less severe symptoms. Since the hospitalization dataset did not specify the severity of Covid-19 symptoms upon admission, a weak supervision machine learning technique was applied to create these two classes of datasets. Weak supervision classification refers to a branch of machine learning which aims to acquire more precisely labelled data for supervised training and modelling. This process is done to inadequately labelled data to acquire a model with good performance [23]. The training of a DNN assisted with describing the clinical and demographic characteristics of patients admitted and was therefore performed on 14 features of patient characteristics (Demographic variables, presence of co-morbidity, care received upon admission and setting of care), to separate the two classes of data sets into: a) severe disease class (as a proxy measure of higher severity, which included those who died during admission or were admitted into an intensive care (ICU) or high care unit (HCU) upon admission), and b) less severe disease class (as a proxy measure, which represented patients admitted to a general ward upon admission or those discharged alive). Both datasets were dis-aggregated by gender. COVID-19 in-hospital mortality was defined as those who faced a COVID-19 related death during their hospital stay and excluded individuals who died from other causes.

The use of the classification method further assisted in uncovering more detailed observations for age and some of the co-morbidity present in the data set. Multi-class classification was done to identify differences in hospitalization frequency, disease severity and mortality among males and females. The multi-class classification further sorted the data into five different classes, each class representing 20%

of the total dataset (a total of five classes). The multi-class classification also provided more detailed analysis of the features used in the model and made it possible to compare disease severity across the different classes, whereas the binary classification merely reported on “severe” and “less severe” case of respective diseases. Additionally, multi-class classification does not have the idea of normal and abnormal outcomes, in contrast to binary classification. Instead, instances are grouped into one of several well-known classes. The process of classifying COVID-19 patients is done due to the varying symptoms and hospital outcomes; the data was grouped accordingly from most severe (Class 1) to least severe (Class 5). A t-test was used to determine the statistical significance in the identified variables which accounted for gender differences in diseases severity. This test was performed on co-morbidities which showed the greatest gendered differences.

Pre-processing

The data-set was cleaned during the pre-processing stage, this included the removal of null values, and incomplete or corrupted data (unusable, unreadable, or inaccessible).

Feature selection

Feature selection in machine learning is the process of selecting the most relevant features/ variables for the machine learning model. This is done to reduce the number of input variables to enhance the performance of the model. The weak supervision feature selection method takes place through a density-based feature categorization in which two similar measures are used for continuous or discrete features. The categorization of patients was done according to their risk and was separated by gender (multi-class classification). Table 1 shows the features selected to feed into the model. Amongst the features, eleven co-morbidities (pre-existing medical conditions) were included for this analysis, namely: HIV, Asthma, cardiac disease, chronic renal failure, chronic pulmonary disease, obesity, diabetes, hypertension, Tuberculosis (TB), past diagnosis of TB, and malignancy.

Table 1 Variables selected for analysis

Category	**Variable**
Demographic Variables	Age Age groups
Setting of care upon admission	ICU High Care General Ward
Care received on admission	Oxygen, not ventilated Oxygen, ventilated No oxygen (room air)
Co-morbidity	Number of co-morbidities Type of co-morbidity

Outcomes	Length of stay before outcome Discharged alive Deceased

Data splits

The data-set was split to allow two subsets, one for training the model and the other for testing the performance of the model. This process enabled cross validation which seeks to achieve stable and confident estimates of how the model performs. For this study the data was split with 70% of the data used for training, 20% for testing and 10% used for validation.

Evaluation metrics

Due to the multidimensional model used, a Receiver Operating Characteristic (ROC) curve was used to measure the overall performance of the classification by the DNN. From the ROC curve, the area under the curve could be calculated, this provided the model's performance score. The higher the score the greater the performance of the DNN and the more accurate the results. The evaluation of the performance of the model is essential for determining the goodness of the model. As part of the intrinsic metrics, the accuracy metric was used to evaluate the performance of the model and determine how accurate the predicted data points were (Appendix I).

Results

Machine Learning Algorithm Output

Results from the DNN output provided essential information needed for understanding the gender differentials. The distribution below is the DNN response of our classification algorithm which classifies between the less severe case of disease and severe. From the distribution we are able to see that there is minimal difference between patients with severe and less severe case of disease. Results from the DNN output and performance testing reported that the model performed well. (Figure 2).

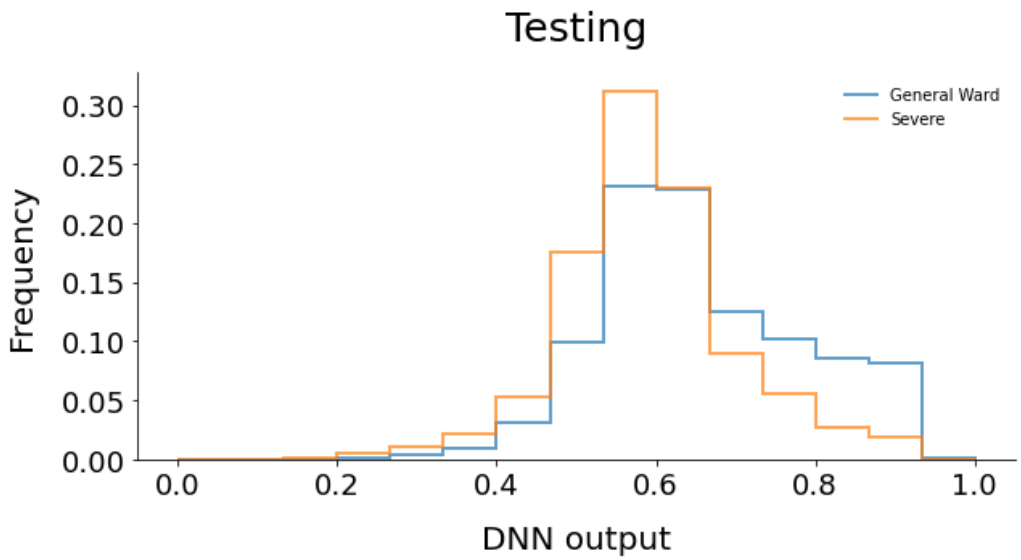

Figure 2 Deep Neural Network output (Frequency normalized to unity)

Hospitalization frequency by gender across 4 waves

A sum of 147,141 patients were admitted to hospital for COVID-19 in the province between 7 March 2020 and 25 March 2022. The number of Covid-19 hospitalisations was highest in wave 3 for both males and females, and higher in females than males across all 4 waves. Whereas the lowest number of COVID-19 hospitalizations occurred during the first wave (Table 2).

Table 2 Sex-disaggregated hospitalization frequency by COVID-19 wave

Epidemic Wave	Date	Number of Hospitalizations total	Female	Male	Unknown	Female/male ratio
Wave 1	7 June 2020 – 22 August 2020	16 886	9 237 (54.7%)	7 589 (44.9%)	40 (0.2%)	1 : 0.8
Wave 2	15 November – 6 February 2021	26 022	13 887 (53.4%)	12 111 (46.5%)	24 (0.09%)	1 : 0.9
Wave 3	9 May 2021 - 18 September 2021	77 659	41 316 (53.2%)	36 264 (46.7%)	79 (0.1%)	1 : 0.9
Wave 4	21 November 2021 – 22 January 2022	19 417	11 216 (57.8%)	8 169 (42.1%)	32 (0.2%)	1 : 0.7

Gender differences in hospitalization frequency by age

Waves one, two and three were analysed and grouped into one category, whereas wave 4 was analysed as is and compared to the first three waves. This grouping was done as similar trends were observed in the first three wave in comparison to the fourth. The observed difference in COVID-19

hospitalization frequency between men and women was highest for wave 4 in the 20 - 40-year age group with a ratio of 1:3 for COVID-19 hospitalisation between men and women, as illustrated in figure 3. Within the severe disease class, similar hospitalization frequencies were observed for all ages across the four waves, except for those aged between 50 and 80 years, where a higher frequency was observed. Similarly, less severe disease showed similar frequencies between male and females, with the exception of the 20–40-year age group, where females had three times more hospital admissions (Figure 3).

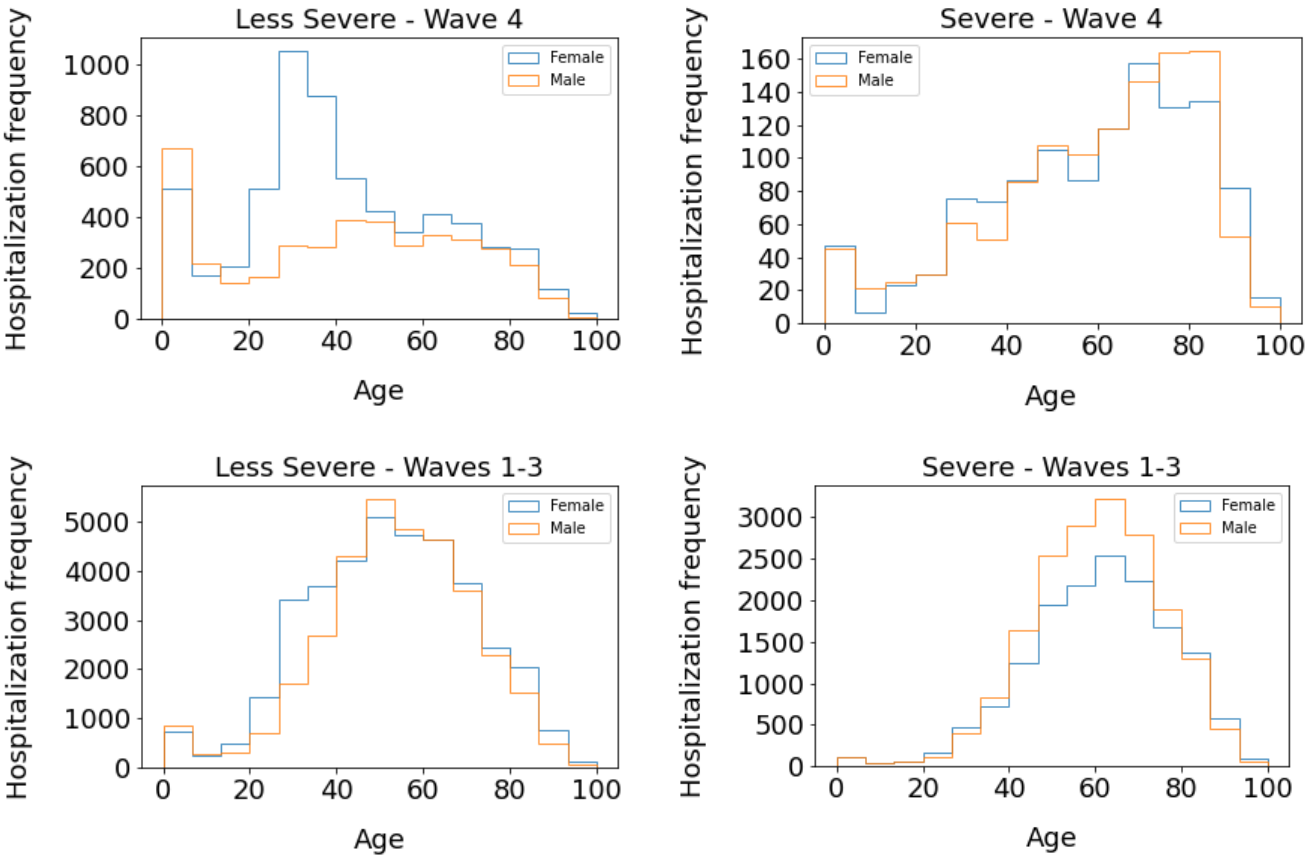

Figure 3 Age distribution of Covid-19 hospitalisations by age and gender for waves 1-3 and wave 4

Gender differences in Covid-19 hospitalisation by co-morbidity

Results from the analysis of co-morbidity uncovered gender differences for hypertension, diabetes, and HIV frequencies across all age groups.

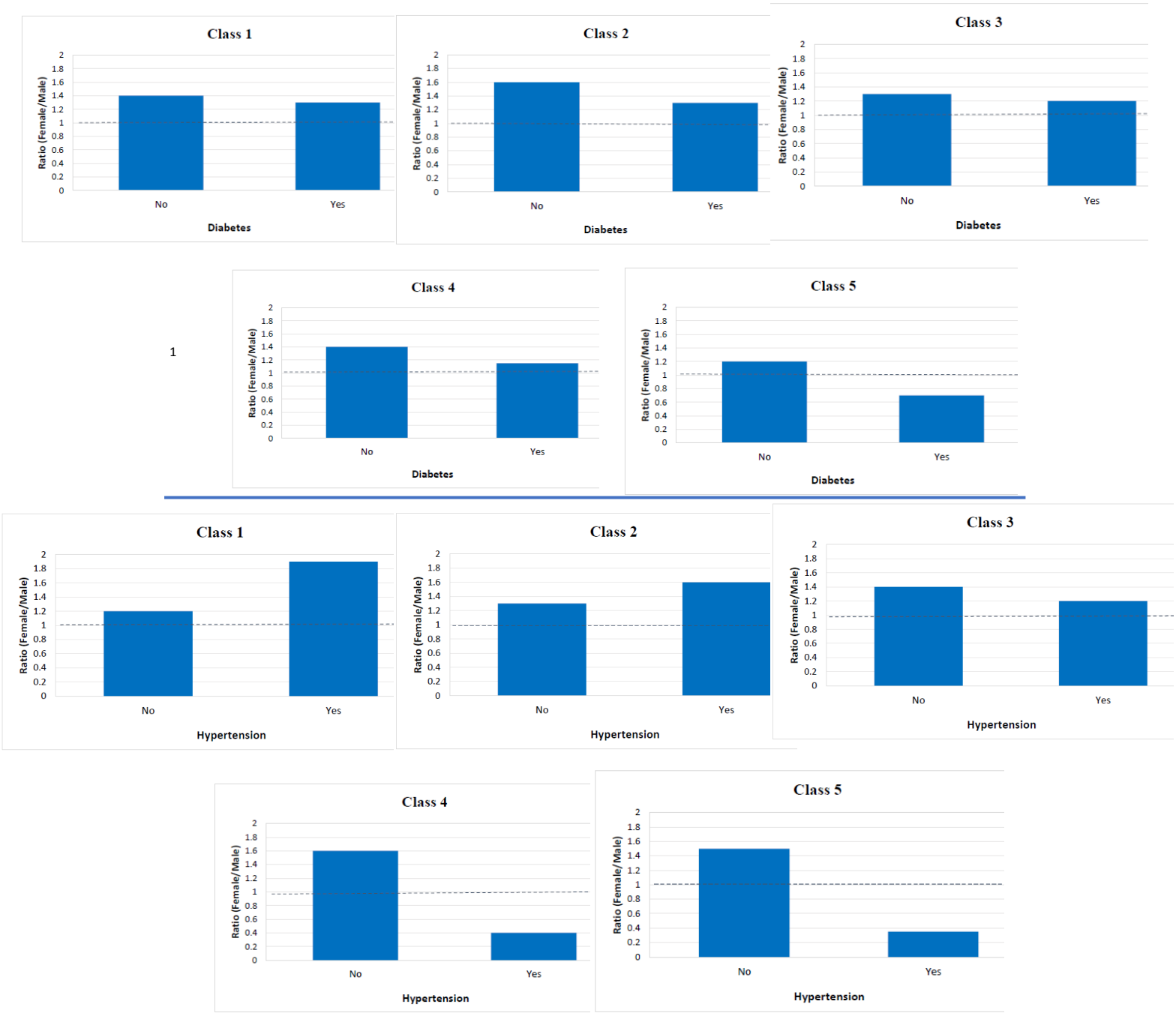

Figure 4 Gender dis-aggregated class graphs (ratio)

[1] class 1 = most severe and class 5 = least severe

Like hypertension, females with diabetes experienced higher frequencies of severe disease in comparison to males for the fourth wave, where there are no noticeable differences in the less severe cases of disease (Figure 4). The use of class graphs represents the differences in hospitalization outcomes between males and females, across the "co-morbidity" dimension used in the model.

Results from the t-test analysis show there were statistically significant differences in hypertension frequencies between the sexes in the first three waves, however showed no significance in the fourth wave. Diabetes showed no significance across all four waves. Lastly HIV showed statistically significant differences between the sexes across all four waves of the pandemic (Table 3). The t-test was performed at a 95% confidence interval. A p-value less than 0,05 was considered as significant, whereas any value greater than 0.05 was considered not significant.

Table 3 Comorbidities in COVID-19 hospitalised patients by gender in South Africa

Period: 17 March 2020 – 15 November 2021 (wave 1-3)			
Comorbid condition	**Male**	**Female**	**P-value***
Hypertension	14 188 (47%)	15 871 (53%)	<0.05
Diabetes mellitus	8 339 (50%)	8 323 (50%)	>0.05
HIV	1 537 (40%)	2 235 (60%)	<0.05
Period: 16 November 2021 – 25 March 2022 (wave 4)			
Hypertension	1 074 (41%)	1 517 (59%)	>0.05
Diabetes mellitus	467 (41%)	665 (59%)	>0.05
HIV	306 (38%)	504 (62%)	<0.05

Hospitalisation frequency by geographical location

Larger percentages of females were admitted in comparison to males across all five districts (data not shown). The data shows that sub-districts containing a larger amount of impoverished/ less affluent suburbs have shown higher female to male ratio in comparison to sub-districts consisting of more affluent suburbs (Figure 5). The city of Tshwane and city of Johannesburg Metropolitan Municipalities have demonstrated a large imbalance in female to male hospitalization frequencies with areas such as Tshwane region 3 and Johannesburg region F both showing ratios of approximately 2:1 for females to males in the third wave. In contrast, Johannesburg region B had a 1:1 ratio for females to males. Similarly, the sub-district of Tshwane region 4 shows a balanced ratio of female to male hospitalizations for the first three waves. Evidently, the fourth wave has demonstrated a larger increase in the gender imbalance across all sub-districts, with three districts having no data for this time period as seen in Figure 5. Differences in hospitalization frequency remained balanced on average between males and females, therefore the data reported in the graph speaks to the fourth wave where gender differences were observed across the sub-districts.

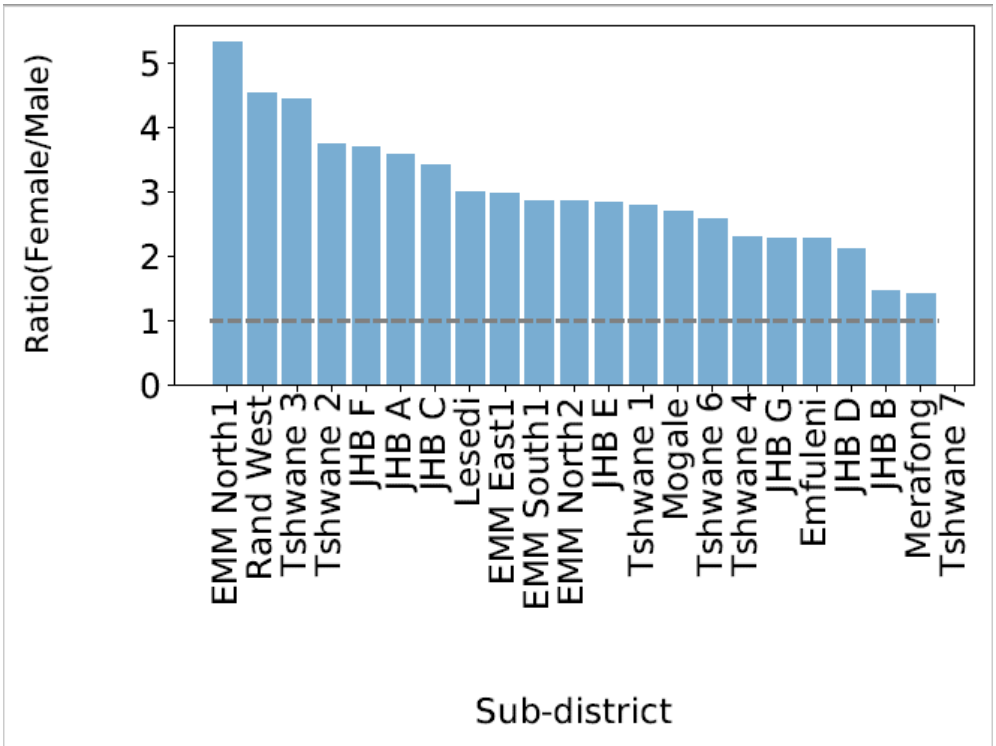

Figure 5 Hospitalisation frequency ratio by geographical location (Gauteng sub-district) for the fourth wave

Discussion

The COVID-19 pandemic has highlighted the impact of gender inequality and the need for policies that address and take into consideration gender differentials in vulnerability and disease outcomes, which occur during disease outbreaks. Despite recent studies highlighting the gender differences in COVID-19 incidence, severity, and outcomes, there is still limited gender-disaggregated COVID-19 data in many countries needed to address existing knowledge gaps. Gender disaggregated data analysis will not only inform deeper understanding of gender vulnerabilities, but also aid in generating evidence-based recommendations for decision makers [8]. The analysis in this paper has demonstrated that multidimensional data analysis using machine learning provides more granular and accurate data on sub-population risk, vulnerabilities, and outcomes of COVID-19.

Our study shows that overall women accounted for 53.5% of hospitalisations, among those who had one or more co-morbidity women accounted for 63%. The observed gender difference in hospitalization was greater in the fourth wave in comparison to the first three waves, a period when

the omicron variant became more prevalent. Results from the t-test show the observed differences were significant. The difference in hospitalization between men and women observed in the fourth wave conveyed that despite females in the 20-to-40-year age group showing higher numbers of hospitalization, these occurred mostly in the general ward of various health facilities within the province, indicating less severe cases of disease in women.

In the context of Gauteng province, our study showed women experienced higher hospitalization frequencies in comparison to males for less severe cases of disease within the 20–40-year age group. Women who suffer from hypertension, diabetes and HIV were more severely affected by COVID-19 than other women without such pre-conditions during the fourth wave of the pandemic, a period in which the omicron virus was the more transmitted variant. Although associations of age, sex and race in COVID-19 patients have been discussed in previous literature [13] , the focus of our study, on the severity of disease by age, co-morbidity and sex, have not extensively been reported on. Both hypertension and diabetes are known to be diseases closely linked to lifestyle and are commonly termed as diseases of modern society due to its association with increased work pressure, a sedentary lifestyle, and medication regimen [4].

Despite, females experiencing less severe cases of disease overall. Those who have the co-morbidity hypertension specifically, have experienced more severe cases of disease, in comparison to men who show a higher frequency in the less severe disease cases for the fourth wave. The gender difference for hypertension and disease severity were higher overall throughout the first three waves of the pandemic. During this period females with hypertension have shown higher frequencies in both severe and less severe cases of disease. Among those who are HIV positive, females have shown higher frequencies of severe disease when compared to males in the same age group, and across the same period. Whereas for the less severe disease case there were no noticeable gender differences observed.

Hypertension, HIV and diabetes are asymptomatic health conditions of a chronic nature therefore the differences between male and women could be explained as a reflection of the differences in health seeking behaviour between both genders [4]. Other possible explanations to the observed differences include socio-demographic and economic determinants of health such as age, gender, marital status, race, education, and geography/provincial location [24]. Other co-morbidity included in the analysis did not show any major gender differences across all four waves of the pandemic, the lack of differences may be an indication of no difference in effects of the different variants in each wave. A recent report by Jassat et al.,[13] stated that the COVID-19 mortality rate among those who are 70 years or younger was approximately 67% which is 12.6 times the expected rate in high income

countries. This difference between high- and lower-income countries in mortality among the younger age groups is partly due to the differences within the population structure, inclusive of age, sex, access to adequate healthcare and co-morbid conditions among other lifestyle factors.

The kinds of NCDs experienced by males and females differ, these differences can be understood by looking at biological variations, which affect the growth of risk factors as well as the manifestation of NCDs. Globally, higher percentages of men have been seen engaging in behavioural risk factors for NCDs in comparison to women [25]. Literature substantiates that the associations between hypertension and diabetes, and hypertension and cardiac disease, are greater in women than men, making hypertension a bigger risk for women, given its association with diabetes and cardiac co-morbidity [26]. Women have additional risk factors for NCDs considering their genetic and hormonal make-up. Within the medical scheme population (across South Africa), the likelihood of being male with hypertension was 1.17 times greater than the likelihood of being female and having the same disease, while the likelihood of being male with diabetes was 1.38 times greater in comparison to being a female with diabetes - the results showed statistical significance and were adjusted for age [27].

Ischaemic heart disease and stroke are also known to be the leading causes of mortality for the female sex and cause the greatest burden of disability. Individuals with lower education levels are more probable to acquire diabetes, however women with low education levels have higher mortality rates from diabetes compared to men of a similar education level [21]. Females generally have a higher life expectancy and therefore are more apt to live longer, they're biologically at higher risk of NCDs. We've also seen that females tend to have higher risk factors as a result of biological sex factors. Therefore, the age and sex standardisation of NCD prevalence rates is essential to forecast the future burden of disease [26].

Health and social vulnerability within townships and informal settlements, are heavily influenced by limited access to medical insurance, pre-existing poor health status, food insecurity and past inability to access healthcare when it was needed [28]. The geography included in this analysis were a mix of affluent and less affluent suburbs. The data portrayed how less affluent sub-districts had a dramatic increase in gendered difference for hospitalization in comparison to more affluent sub-districts. This observation however was not the same throughout the fourth wave of the pandemic, as all sub-districts faced increased differences in hospitalizations as seen in Figure 5. The link between geography, age and poor health outcomes for COVID-19 can be clearly seen and the need for evidence-based solutions to address the association is paramount. Our study findings are similar to that of Jassat et al., 2022 who made use of traditional analytical methods and found sex differences in the

mortality of those 40 years and younger. Additionally, Finn et al., utilized similar machine learning techniques to develop an early detection system which predicted daily COVID-19 cases. The development of this system was beneficial for predicting the start of the third wave of COVID-19 and inform public health policies [19]. The use of AI in healthcare enhances the research process by increasing data flow which further improves predictability of disease for large amounts of data in a more efficient manner that traditional methodologies cannot. AI furthermore improves data analysis by managing large volumes of data and performing big data analytics. The use of AI concepts such as machine learning can potentially transform healthcare by outperforming not only standard epidemiological methods of analysis but also outperforming clinical systems and modelling complex correlations between hidden factors of data [29].

Conclusion

In essence, this study adds to the evidence of gender differences in COVID-19 hospitalization in the Gauteng province. The study found that overall, men who experienced hospitalization due to COVID-19 were more probable to have more severe disease as well as death in comparison to women. However, women who were hospitalised for covid-19 were more probable to have “less severe disease” and suffer from HIV, hypertension, and diabetes, dependent on age, location, and wealth status. Therefore, public health efforts should focus on addressing lifestyle factors which contribute to differences in co-morbidity between men and women in the country. The findings of this study should be used to inform policy and target government programs to those who are most at risk of future infectious diseases due to their existing co-morbidity, geographical location, age, and sex. This study demonstrated the critical importance of machine learning using multidimensional data analysis of dis-aggregated data by sex to provide more accurate, real-time information to inform more precise public health monitoring, planning and responses that support more inclusive and equitable health for all.

References

[1] Naidu T. The COVID-19 pandemic in South Africa. Psychological Trauma: Theory, Research, Practice, and Policy. 2020;12(5):559.

[2] Fund UNP. COVID-19: A gender lens: Protecting sexual and reproductive health and rights and promoting gender equality. 2020.

[3] Bischof E, Oertelt-Prigione S, Morgan R, Klein SL, Sex, in COVID19 Clinical Trials Working Group (SGC) G, et al.. Towards precision medicine: inclusion of sex and gender aspects in COVID-19 clinical studies—acting now before it is too late—a joint call for action. MDPI; 2020.

[4] Fatouros S, Capetola T. Examining Gendered Expectations on Women's Vulnerability to Natural Hazards in Low to Middle Income Countries: A critical Literature Review. International Journal of Disaster Risk Reduction. 2021;64:102495.16

[5] Peckham H, de Gruijter NM, Raine C, Radziszewska A, Ciurtin C, Wedderburn LR, et al. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. Nature communications. 2020;11(1):1–10.

[6] Alavinejad M, Mellado B, Asgary A, Mbada M, Mathaha T, Lieberman B, et al. Management of Healthcare Resources in the Gauteng Province, South Africa, During the COVID-19 Pandemic. South Africa, During the COVID-19 Pandemic (March 3, 2022). 2022.

[7] Ekpanyaskul C, Padungtod C. Occupational health problems and lifestyle changes among novice working-from-home workers amid the COVID-19 pandemic. Safety and health at work. 2021;12(3):384–389.

[8] Wan KS, Tok PSK, Yoga Ratnam KK, Aziz N, Isahak M, Ahmad Zaki R, et al. Implementation of a COVID-19 surveillance programme for healthcare workers in a teaching hospital in an upper-middle-income country. PloS one. 2021;16(4):e0249394.

[9] Ya'qoub L, Elgendy IY, Pepine CJ. Sex and gender differences in COVID-19: More to be learned! American Heart Journal Plus: Cardiology Research and Practice. 2021;3:100011.

[10] Korkoyah Jr DT, Wreh FF. Ebola impact revealed: An assessment of the differing impact of the outbreak on the women and men in Liberia. 2015.

[11] Casale D, Shepherd D. The gendered effects of the COVID-19 crisis and ongoing lockdown in South Africa: Evidence from NIDS-CRAM Waves 1-3. National Income Dynamics Study Coronavirus Rapid Mobile Survey (NIDS-CRAM) Working Paper. 2021.

[12] Gauteng hospitalisation: Covid-19 South Africa;. Available from: https://www.covid19sa.org/gautenghospitalisation.

[13] Jassat W, Ozougwu L, Munshi S, Mudara C, Vika C, Arendse T, et al. The intersection of age, sex,

race and socioeconomic status in COVID-19 hospital admissions and deaths in South Africa. South African Journal of Science. 2022;118(5-6):1–14.

[14] Abdullah F, Myers J, Basu D, Tintinger G, Ueckermann V, Mathebula M, et al. Decreased severity of disease during the first global omicron variant covid-19 outbreak in a large hospital in tshwane, south africa. International Journal of Infectious Diseases. 2022;116:38–42.

[15] Yu D. Revisiting the COVID-19 vulnerability index in South Africa. Development Southern Africa.2021:1–18.

[16] Shifa M, David A, Leibbrandt M. Spatial inequality through the prism of a pandemic: Covid-19 in South Africa. Scientific African. 2021;13:e00949.

[17] Mellado B, Wu J, Kong JD, Bragazzi NL, Asgary A, Kawonga M, et al. Leveraging artificial intelligence and big data to optimize COVID-19 clinical public health and vaccination roll-out strategies in Africa. International Journal of Environmental Research and Public Health. 2021;18(15):7890.

[18] Muhammad L, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset.SN computer science. 2021;2(1):1–13.

[19] Stevenson F, Hayasi K, Bragazzi NL, Kong JD, Asgary A, Lieberman B, et al. Development of an early alert system for an additional wave of covid-19 cases using a recurrent neural network with long short-term memory. International Journal of Environmental Research and Public Health.2021;18(14):7376.

[20] The Power of Collaboration, Artificial Intelligence and Big Data in the fight against COVID-19 in Africa Available from: http://acadic.org/

[21] World Health Organization, et al. Why using a gender approach can accelerate noncommunicable disease prevention and control in the WHO European Region: WHO European high-level conference on noncommunicable diseases: time to deliver–meeting NCD targets to achieve Sustainable Development Goals in Europe: 9–10 April 2019, Ashgabat, Turkmenistan. World Health Organization.Regional Office for Europe; 2019.

[22] National COVID-19 daily report - NICD;. Available from: https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/surveillance-reports/national-covid-19-daily-report/.

[23] Zhou ZH. A brief introduction to weakly supervised learning. National science review. 2018;5(1):44–53.

[24] Abdul-Razak S, Daher AM, Ramli AS, Ariffin F, Mazapuspavina MY, Ambigga KS, et al. Prevalence, awareness, treatment, control and socio demographic determinants of hypertension in Malaysian adults. BMC public health. 2016;16(1):1–10.

[25] World Health Organization. Gender and noncommunicable diseases in Europe: analysis of STEPS data. World Health Organization Regional Office for Europe; 2020.

[26] Di Giosia P, Giorgini P, Stamerra CA, Petrarca M, Ferri C, Sahebkar A. Gender differences in epidemiology, pathophysiology, and treatment of hypertension. Current atherosclerosis reports.2018;20(3):1–7.

[27] Consultants PA. Brief Three: Sex, gender and NCDs; 2022. Available from: https://percept.co.za/2022/03/11/brief-three/.

[28] De Kadt J, Gotz G, Hamann C, Maree G, Parker A. Mapping vulnerability to COVID-19 in Gauteng. GCRO Map of the Month, Gauteng City-Region Observatory. 2020.

[29] Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database. 2020 Jan 1;2020.

Appendices

Appendix I: Evaluation Metric – Roc Curve

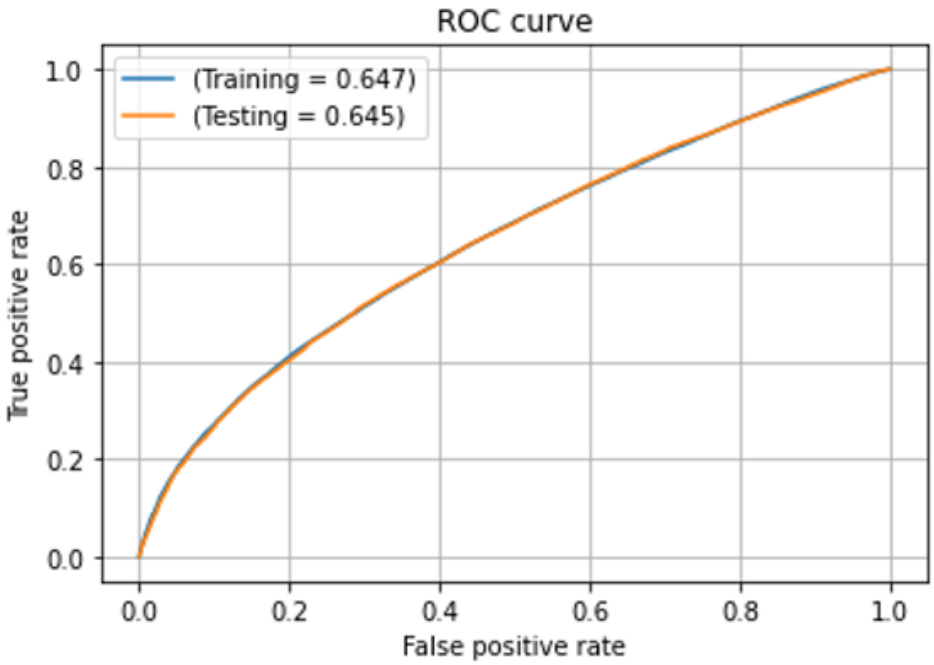

Figure 3 Roc curve

Appendix II: Ethics Approval

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Miss Meghan Malaatjie

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M2111127

NAME: (Principal Investigator)	Miss Meghan Malaatjie
DEPARTMENT:	School of Public Health
PROJECT TITLE:	The use of machine learning techniques in identifying gender differentials in COVID-19 hospitalizations, probabilities of hospitalization outcomes and hidden correlations with demographic and clinical factors
DATE CONSIDERED:	26/11/2021
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Dr Mary Kawonga and Mr Bruce Mellado
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	27/02/2023

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions.Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report.** The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **November** and will therefore be due in the month of **November** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

4. MANUSCRIPT DESCRIPTION AND REQUIREMENTS BY ARTICLE CATEGORY

"A. Original Research.

The journal welcomes clinical trials, observational studies, studies of diagnostic tests, cost-effectiveness analyses, systematic reviews and meta-analyses, and qualitative research, among other types of submissions.

The journal endorses the use of an appropriate reporting guideline when writing any health research manuscript. You can find the most commonly required reporting guidelines on the **EQUATOR Network**, which also gives general information on how to choose the correct guideline. The EQUATOR Network collects more than 370 reporting guidelines for many study types, and authors are encouraged to follow the appropriate ones for their study. At minimum, your article should report the content addressed by each item of the identified checklist or state that the item was not considered in the study (for example, if you did not use blinding, your article should say so).

Manuscripts must be relevant to an international audience, and must be scientifically and technically sound, with appropriate methodology and presentation, and reaching conclusions adequately supported by the data. Additional details about our acceptance criteria, can be found **here**.

Manuscripts should conform to the "**Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly work in Medical Journals**," developed by the International Committee of Medical Journal Editors (ICMJE), and include an appropriate Introduction, Methods, Results, and Discussion sections.

Regarding methods, manuscripts should include an appropriate description of the selection and description of the participants (if appropriate), the dates and setting(s) of the study, data sources used, study design, interventions/exposures, diseases/outcomes, covariates considered and included, participant follow-up (if any), materials, and data analysis and statistical approaches employed.

i. Statistical Methods and Data Presentation

As recommended by the ICMJE, please describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to judge its appropriateness for the study, and to verify the reported results.

Authors are encouraged to review and follow the recommendations put forward in the "**Guidelines for reporting of statistics for clinical research in urology**" (Assel et al., 2018) for guidance on the proper analysis, reporting, and interpretation of clinical research.

When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Use means and standard deviations (SDs) for normally distributed data, and medians and ranges or interquartile ranges (IQRs) for data that are not normally distributed.

Whenever possible, proportions and percentages should be accompanied by the actual numerator and denominator from which they were derived.

Avoid relying solely on statistical hypothesis testing, such as P values, which fail to convey important information about effect size and precision of estimates. P values should never be presented alone without the data that are being compared and the test used to derive them. If P values are reported, please follow standard conventions for decimal places: for P values less than .001, report "P<.001"; for P values between .001 and .01, report the value to the nearest thousandth; for P values greater than or equal to .01, report the value to the nearest hundredth; and for P values greater than .99, report as "P>.99."

References for the design of the study and statistical methods should be to standard works when possible. Define statistical terms, abbreviations, and symbols. Further, distinguish prespecified from exploratory analyses, including subgroup analyses.

Do not use terms such as "marginal significance" or "trend toward significance" for results that are not statistically significant at the defined threshold.

At the end of the Methods section, please describe all of the statistical tests used for the analyses. State any a priori levels of significance, and whether tests were 1- or 2-sided. Also, specify the statistical software package(s) used in the analyses, and its versions. We encourage authors to follow **SAMPL guidelines**.

When constructing descriptive multivariable models, univariable variable selection, that is including those independent variables in a multivariable model that show significant association with the outcome in univariable models, **is strongly discouraged**. Such univariable prefiltering (i.e. selecting variables for the multivariable model based on those with a univariable p-value of less than a defined threshold) does not add stability to the selection process as it is based on stochastic quantities, and can lead to overlooking important adjustment variables needed for control. Authors are, instead, encouraged to review alternative available variable selection methods that are based on significance or information criteria, penalized likelihood, the change-in-estimate criterion, background knowledge, or combinations thereof, summarized by **Heinze et al. (2018)**.

With regards to **observational causal inference studies,** a Working Group of editors of respiratory, critical care, and sleep journals has **published guidance** on appropriate and inappropriate analytic methods for such studies, including cohort, case-control, and cross-sectional studies. We encourage authors to familiarize with that document and follow its recommendations. In addition, we concur with the criteria described in the **Author**

Guidelines of the Annals of the American Thoracic Society (AnnalsATS), which states that authors should keep in mind that:

1. Papers reporting the results of observational causal inference studies should include a detailed plan for how potential confounding factors were selected and handled. Confounders should be defined based on mechanistic and biological knowledge.

2. Covariate selection should not be based on statistical hypothesis testing, changes in effect estimates, model fit, or any stepwise methods. P-value– or model-based variable selection methods, variable selection based on beta-coefficient changes, and selection of variables to identify multiple "independent predictors", are strongly discouraged.

Similarly, as stated in the a forementioned guidelines, when constructing **prediction models** "modern variable selection methods are favored. These include (but are not limited to) shrinkage, penalized maximum likelihood estimation, and machine learning. P-value–based selection methods (e.g., stepwise selection) are discouraged (unless the paper includes validation of the prediction model in an external cohort). Authors should ensure that they address basic prediction model issues such as overfitting, optimism, missingness, interactions, model assumptions, and linearity". For additional information, please consult Steyerberg's "Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating" (2019). For prediction models, authors are encouraged to follow the **TRIPOD statement** and the guidelines put forward in the "**Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals**", by Leisman et al (2020).

ii. Clinical Trials

Clinical trials should follow the CONSORT guidelines and authors will be asked to provide the **completed associated checklist** upon submission. The report must include a figure detailing information about the progress through the phases of the trial, including enrollment, patient allocation, follow-up, and analysis.

B. Literature (Narrative) Review

Literature reviews should provide an authoritative, timely, and evidence-based synthesis of the literature on a specific topic of interest to the health sciences community. Reviews should be focused and provide an update on our current understanding of a specific condition, a diagnostic consideration, or treatment. While unsolicited proposals for review articles are welcome, review articles are most commonly invited, so authors interested in submitting a review are encouraged to contact the editorial office before submitting.

C. Study Protocol

The journal welcomes articles that report the design and methods of clinical trials or systematic reviews and meta-analyses.

i. Clinical Trials

Authors must follow **SPIRIT guidelines** or associated extensions.

ii. Systematic Review and Meta-Analysis

Authors must follow **PRISMA-P guidelines**.

D. Methods

The journal welcomes articles focused on methodological approaches to healthcare research.

Articles in this category include but are not limited to those describing the development and testing of new laboratory assays and methods (with an emphasis on clinical laboratory testing), articles on the development or testing of novel approaches for systematic reviews/meta-analyses and clinical trials, and studies on methodology or the development of tools for quantitative and qualitative research. In addition, the journal welcomes articles describing methodological approaches for the design and analysis of epidemiologic studies.

Given the high degree of technical expertise that is sometimes required to evaluate these contributions, we ask authors to please write their methodological manuscripts in a way accessible to clinicians and other health science researchers and include all non-essential highly technical material in the supplementary information.

E. Editorial

These are written by the Chief Editors or are specifically commissioned, and usually discuss issues pertaining the journal and its operation or provide a forum to introduce special issues or collections. Editorials are not externally peer reviewed.

F. Perspective

These are authoritative and expert opinion pieces involving areas of broad interest. These are typically invited, but we may consider unsolicited pieces of significant interest. Perspectives may address virtually any important topic in the health sciences. Perspectives should not be used as short review articles, and instead, should put forward the authors' opinion on a specific subject.

G. Commentary

Commentaries are short comments by experts on the field on an article recently published in the journal. Commentaries should state the advance being discussed and provide a critical evaluation of the research. In addition, the article should provide some background information and explanation that allows the placing of the findings in the right context. Including some questions to be addressed in future research, is encouraged.

These pieces are usually short (usually less than 1200 words) and are aimed at a non-specialist audience.

H. Correspondence

Correspondence pieces succinctly comment on an article recently published in the journal. Letters should be focused and short (< 1500 words), and include no more than one display item. Correspondence being considered for publication will typically be sent to the authors of the original article, who will be given the opportunity to reply.

Correspondence discussing a topic unlinked to an article in the journal but briefly addressing a relevant topic for the community, might also be considered. Such articles are not externally peer reviewed.

Correspondence must not duplicate other material published or submitted for publication and should not include unpublished data.

I. Guidance Paper

Guidance articles are public statements of what a representative group of experts agree to be evidence-based and state-of-the-art knowledge on an aspect of practice/policy. They should not exceed 1500 words.

J. Observations

Observation articles are detailed reports or presentations of the symptoms, signs, diagnosis, treatment, and follow-up of an individual patient or selected disease indications. They usually describe an unusual or novel occurrence or have substantial learning value for readers. They should not exceed 3000 words.

K. Letter to the Editor

Letters should encourage scientific discussion on topics pertaining to health science, and may also consist of short reports or observations. If discussing a recent publication, it should be no more than 4 weeks since the article's publication. A Letter to the Editor should start with 'Editor,' and be short and to the point. Long introductions and discussions are discouraged. Letters should have title page, main text, no abstract, no key words, and no subheadings. Letters should not exceed 1000 words and should have a maximum of 5 references and 1 table or 1 illustration. Letters should have no more than 3 authors. The editors reserve the right to edit letters for clarity and brevity.

J. Research Letter

Research letters succinctly present the findings of studies using a simple design and answering a single, focused, and specific question, or those of small-scale, early-stage studies. The latter may include work in progress or preliminary studies with a small sample size that call for further investigation. Such exploratory studies need to include a clear discussion of the limitations and should ideally be presented as hypothesis-generating rather than confirmatory.

Letters have a 1200 word limit (excluding references, tables, and figures), and typically have fewer than 20 references and no more than 1-2 display elements (figures/tables) in total. Letters do not have an abstract.

5. MANUSCRIPT PREPARATION: GENERAL INSTRUCTIONS

Manuscripts must be written in English. Authors interested in language editing, translation, article formatting, and figure preparation services can contact **Wiley Editing Services**. In addition, Wiley has a range of resources for authors preparing articles for submission, available **here**. In particular, authors may benefit from referring to Wiley's best practice tips on **Writing for Search Engine Optimization**.

Nonstandard abbreviations should not be used. In general, terms should not be abbreviated unless they are used repeatedly and the abbreviation is helpful to the reader. Initially use the word in full, followed by the abbreviation in parentheses. Thereafter, use the abbreviation only. Specialized jargon should be avoided in favor of widely recognized scientific and clinical language.

Measurements should be given in SI or SI-derived units.

Chemical substances should be referred to by the generic name only. Trade names should not be used. Drugs should be referred to by their generic names. If proprietary drugs have been used in the study, refer to these by their generic name, mentioning the proprietary name, and the name and location of the manufacturer, in parentheses.

Sequence variants should be described in the text and tables using both DNA and protein designations whenever appropriate. Sequence variant nomenclature must follow the current HGVS guidelines; see **http://varnomen.hgvs.org/**, where examples of acceptable nomenclature are provided.

You will be asked to provide the full address information for the **corresponding author** (see below). Please be sure to do this, as the processing of your article may be delayed without complete address information for the corresponding author.

While articles can, in principle, be of any length, and there is no restriction on word count or number of figures, authors are encouraged to present their findings concisely, and take into consideration typical lengths of their respective article type (see above).

Reporting Guidelines and Methodology:

Please note that manuscripts submitted to the journal must follow the **relevant reporting standards on the EQUATOR network website** Authors are strongly encouraged to use the checklists proposed by these guidelines, which may be requested during review. A summary of the reporting guidelines for main study types is below:

Guidelines	Study Type
CONSORT	Randomized trials

STROBE	Observational studies
PRISMA	Systematic review and meta-analyses
PRISMA-P	Systematic review and meta-analysis protocols
SPIRIT	Study protocols
STREGA	Genetic association studies
SRQR	Qualitative studies
STARD	Diagnostic accuracy studies
AGREE	Clinical practice guidelines
ARRIVE	Animal pre-clinical studies
SRQR	Qualitative research
MOOSE	Meta-analyses of observational studies
STREGA	Gene-disease association studies
CHEERS	Health economic evaluations
TRIPOD	Studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes.

Authors are encouraged to review and follow the recommendations put forward in the "**Guidelines for reporting of statistics for clinical research in urology**" (Assel et al., 2018) for guidance on the proper analysis, reporting, and interpretation of clinical research.

A. Main article file

Articles should typically be organized into sections under headings in the following order: title; the full names of the authors, their institutional affiliations, and contribution to the work; acknowledgements and funding; conflicts of interest statement; abstract and keywords; main text; references; figure legends; tables; figures; appendices (if relevant).

i. Title

The title should be short and non-declarative, and contain the major key words. Titles must include the study design, when appropriate, for example “The trend of top five types of poisonings in hospitalized patients based on ICD-10 in the northeast of Iran during 2012–2018: A cross-sectional study” The title should not contain any abbreviations. A short running title (< 40 characters) should also be provided.

See here for **Wiley’s best practice SEO tips**.

ii. Abstract (where required) and keywords

The Abstract should be divided into the following sections 'Background and Aims', 'Methods’. ‘Results', and 'Conclusion', and it should not exceed 300 words. The Background and Aims section should include the rationale for the study. In the methods, please include dates,

interventions, outcomes assessed, and the statistical approach employed. For Results, please include the most important findings of the study (please be quantitative). For the conclusions, please state conclusions adequately supported by the results and study design, and a sentence about the implications of the study. **If appropriate, please include the name of the trial register and your clinical trial registration number at the end of your abstract.**

Please include 4 – 6 keywords. Keywords can be obtained from MeSH (Medical Subject Headings). The **MeSH browser** provides an online guide to the selection of key words.

iii. Authors

For each author, please include their full name and ORCID ID (if available), and the name of the department(s) and institution(s) to which each author should be attributed. Please note that the submitting author is *required* to provide and ORCID ID. More information about ORCID IDs can be found **here**.

The ICMJE recommends that authorship should be based on the following four criteria:

- Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
- Drafting the work or revising it critically for important intellectual content; AND
- Final approval of the version to be published; AND
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

All those designated as authors should meet all four ICMJE criteria for authorship, and all who meet the four criteria should be identified as authors. These authorship criteria are intended to reserve the status of authorship for those who deserve credit and can take responsibility for the work. The criteria are not intended for use as a means to disqualify colleagues from authorship who otherwise meet authorship criteria by denying them the opportunity to meet criterion #s 2 or 3. Therefore, all individuals who meet the first criterion should have the opportunity to participate in the review, drafting, and final approval of the manuscript.

Prior to submitting the article, all authors should agree on the order in which their names will be listed in the article. **It is the responsibility of the submitting author to obtain such approval by all authors before submission**.

Any change in authorship following the original submission must be justified and agreed to in writing by the affected author(s)

The journal will not publish articles written in part or in full by individuals not named as authors.

The corresponding author should be identified as such. As stated in guidelines of the ICMJE, the corresponding author is the one individual who takes primary responsibility for communication with the journal during the manuscript submission, peer review, and

publication process, and typically ensures that all the journal's administrative requirements, such as providing details of authorship, ethics committee approval, clinical trial registration documentation, and gathering conflict of interest forms and statements (confirming that they are accurate and up-to-date), are properly completed, although these duties may be delegated to one or more coauthors. The corresponding author should be available throughout the submission and peer review process to respond to editorial queries in a timely way and **should be available after publication to respond to critiques of the work and cooperate with any requests from the journal for data or additional information** should questions about the paper arise after publication. Whilst we do not limit the number of corresponding authors, it is at the discretion of the editorial team to request additional supporting documents should there be any doubt as to the extent of corresponding authors' involvement.

Author Contributions

For all articles, the journal mandates the CRediT (Contribution Roles Taxonomy)—more information is available on our **Author Services** site. The contribution statement will be published with the final article and should accurately reflect contributions to the work.

iv. Acknowledgments and funding

Contributions from anyone who does not meet the criteria for authorship should be listed, with permission from the contributor, in the Acknowledgments section.

Authors should also list all funding sources in the Acknowledgments section. It is the responsibility of the corresponding author to review the Acknowledgment section and funding details with all authors to ensure it is accurate. If in doubt, please check the Open Funder Registry for the correct nomenclature: **http://www.crossref.org/services/funder-registry/**.

When listing the funding sources and/or financial relationships, please provide explanations of the role of those sources, if any, in study design; collection, analysis, and interpretation of data; writing of the report; the decision to submit the report for publication. Alternatively, please provide a statement declaring that the supporting source/financial relationships had no such involvement.

v. Conflict of Interest Statement

Health Science Reports requires that all authors disclose any potential sources of conflict of interest and all articles published in *Health Science Reports* must contain a conflict of interest statement, even if this simply states that no conflicts of interest exist for any of the article's authors. During the submission process you will be asked to supply a conflict of interest statement; if you need more details about this, please see the "Conflicts of Interest" section of **Wiley's Best Practice Guidelines on Publishing Ethics**, or ask the *Health Science Reports* editorial team for guidance (**healthsciencereports@wiley.com**). As stated above, it is the responsibility of the corresponding author to review the conflict of interest statement with all authors and collectively to disclose with the submission all pertinent commercial and other relationships.

Any interest or relationship, financial or otherwise that might be perceived as influencing an author's objectivity is considered a potential source of conflict of interest. Potential sources of conflict of interest include, but are not limited to, patent or stock ownership, membership of a company board of directors, membership of an advisory board or committee for a company, and consultancy for or receipt of speaker's fees from a company. The existence of a conflict of interest does not preclude publication.

Authors are encouraged to use the **ICMJE Conflict of Interest form** for preparing their conflicts of interest statements.

If you are not sure whether or not you have a conflict of interest, please ask our editorial team for advice: **healthsciencereports@wiley.com**.

vi. Data Availability Statement

We encourage authors to share the data supporting the findings of their study and the associated code and materials, whenever possible. Examples include individual-level deidentified patient data, survey results, interview transcripts, statistical code, and others.

On submission, we require authors to select a data sharing statement in their manuscript, indicating whether data and materials are available, and, if so, how and where to access them. For data that cannot be shared, a short description of the restriction will need to be provided.

Health Science Reports reserves the right to request at any time confidential access to any primary data needed to reproduce the findings of the article, so that the results reported can be verified. Editors may also use data availability statements to inform their editorial decisions.

Review **Wiley's Data Sharing policy** where you will be able to see and select the data availability statement that is right for your submission. Data should not be shared in any way that could compromise participant anonymity or privacy, and data sharing should comply with all local laws and regulations, and any licensing agreements.

Where a research community in a particular field has established standards for what, where, and how data should be shared, we expect authors to meet those.

In accordance to ICMJE guidelines, manuscripts submitted to *Health Science Reports* that report the results of clinical trials *must* contain a **data sharing statement.** Data sharing statements must indicate the following: whether individual deidentified participant data (including data dictionaries) will be shared; what data in particular will be shared; whether additional, related documents will be available (e.g., study protocol, statistical analysis plan, etc.); when the data will become available and for how long; by what access criteria data will be shared (including with whom, for what types of analyses, and by what mechanism).

Illustrative examples of data sharing statements that would meet these requirements can be found **here**. In addition, the ICMJE requires that clinical trials that begin enrolling participants on or after 1 January 2019 include a data sharing plan in the trial's registration.

vii. Main text

The journal uses British/US spelling; however, authors may submit using either option as spelling of accepted articles is converted during the production process.

Footnotes to the text are not allowed and any such material should be incorporated into the text in an appropriate way.

viii. References

As with the main body of text, the completeness and content of your reference list is more important than the format chosen. A clear and consistent, generic style will assist the accuracy of our production processes and produce the highest quality published work, but it is not necessary to try to replicate the journal's own style, which is applied during the production process. If you use bibliographic software to generate your reference list (e.g. Endnote), select a standard output style, and check that it produces full and comprehensive reference listings. We recommend using a style that uses a Name-Year system (i.e. CSE system), as this facilitates the work of the reviewers. A guide to the minimum elements required for successful reference linking appears below. The final journal output will use the 'Harvard' style of reference citation. If your article has already been prepared using the 'Vancouver' system, we are quite happy to receive it in this form. We will perform the conversion from one system to the other during the production process.

Example:

Journal Article

Terzolo M, Pia A, Ali A, et al. Adrenal incidentaloma: a new cause of the metabolic syndrome? *Clin Endocrinol Metab*. 2002;87:998-1003.

Book

Council for International Organizations of Medical Sciences. *International Ethical Guidelines for Health-Related Research Involving Humans*. Fourth ed. Geneva, Switzerland, CIOMS; 2016.

References to online research articles should always include a DOI, where available. When referring to other Web pages, please include a date on which the resource was accessed. Authors are responsible for ensuring that the information in each reference is complete and accurate.

If authors cite articles they have participate in that are currently under review at other journals or unpublished, and are deemed important to assess the suitability of the article under consideration at *Health Science Reports*, authors must provide copies of such articles upon request by the editors.

ix. Figure Legends

Legends should be concise but comprehensive – the figure and its legend must be understandable without reference to the text. Include definitions of any symbols used and define/explain all abbreviations and units of measurement. Scale bars should be included in all images for which magnification and scale are relevant.

x. Tables

Tables should be self-contained, and they should complement but not duplicate information contained in the text. They should be supplied as editable files, not pasted as images. Legends should be concise but comprehensive – the table, legend, and footnotes must be understandable without reference to the text. All abbreviations must be defined in footnotes. Footnote symbols: †, ‡, §, ¶, should be used (in that order). Statistical measures such as SD or SEM should be identified in the headings.

Please do not use *, **, *** to refer to specific P-value thresholds. The exact P-values should be presented, when possible.

xi. Figures

Although we encourage authors to send us the highest-quality figures possible for the final version of the manuscript, for peer-review purposes, we are happy to accept a wide variety of formats, sizes, and resolutions.

Color figures: It is preferable that line figures (e.g. graphs and charts) are supplied in black and white so that they are legible if printed by a reader in black and white.

xii. Appendices

Appendices will be published after the references. For submission they should be supplied as separate files but referred to in the text."

Appendix IV: Approved Research Protocol

Contents

Introduction 44

Background 44
Literature review 45
Problem statement: 47
Justification of the research 48
Aim & objectives 48
Aim: 48
Objectives are to: 48
Methods 49
Research design and scope 49
Data source 49
Variables 50
Overview of Proposed Solutions 51
Pre-processing 51
Feature selection 51
Data splits 52
Data analysis 52
Performance testing and optimization 53
Performance evaluation 53
Environmental setup: 54
Research limitations 54
Ethical considerations 55
Project timeline 55
References 55

Introduction

Background

The COVID-19 pandemic has enhanced the health, social, and economic impacts of gender. Since the beginning of the COVID-19 pandemic, women have been bearing the brunt of both social and economic consequences as a result of the pandemic (1). Observed differences in health risks and outcomes between men and women have been reported for numerous diseases inclusive of COVID-19 (2). The COVID-19 pandemic has led to the breakdown of social infrastructure which worsen existing gender vulnerabilities and is thought to play a role in differences in COVID-19 outcomes between men and women. An example of this, a recent study by Mittal & Singh (3) conveyed that existing gender inequality is worsened because of the increase in women's exposure to harassment and violence when attempting to procure essential resources such as food and water (4). A study on COVID-19 by Forrester et al (5), found that men not only less frequently presented for gender neutral inflammatory diseases, but women also experienced longer hospitalization in comparison to men. Similarly, data from the South African National Department of Health conveys that between April 2020 and July 2021, there have been observed differences in hospitalization by gender with women having roughly 5.6% more cases of COVID-19 hospitalization in comparison to men with roughly 47% of the total COVID-19 hospitalization being men and 53% women. The data shows that gender imbalance by hospitalization is a present problem within the Gauteng province (6).

This burden has drastically increased since the beginning of the pandemic. The aggravation of these gender inequalities, which are likely a reflection of gendered social norms, occurred at a time when many countries were in a lockdown, and women had increased caregiving responsibilities (2).

Gender plays an essential role in shaping health inequalities (e.g., inequalities between men and women in health access or health-seeking behavior), which further influence unequal health outcomes between men and women [7,8]. Some studies have reported lower incidence of COVID-19 infections in women with others reporting lower COVID-19 infections in men (8) and some reporting no difference at all (9,10). The reason for the differences or lack thereof are not entirely understood, as

it could be an intersection of multiple factors, including gender-based vulnerabilities (11). Additionally, many asymptomatic cases in addition to missed diagnosis each contribute to knowledge gaps in regard to COVID-19 incidence. Despite recent studies highlighting the sex and gender-based differences in COVID-19 incidence, severity, and outcomes (12), there is still limited gender-disaggregated COVID-19 data in many countries to address existing knowledge gaps. Gender disaggregated data analysis will not only inform deeper understanding of gender vulnerabilities, but also aid in generating evidence-based recommendations for decision makers (13).

Literature review

Early data has indicated that among the paramount factors which form part of the determinants that influence the susceptibility and vulnerability to COVID-19, gender and sex not only intersect with each other but also with other social determinants (14). Gender has been defined as the socially constructed collection of roles, relationships, behaviors, and influence which society assigns to the two different genders. Gender most commonly speaks to the societal-defined differences, whereas sex refers to the biological differences between males and females (15). For the purpose of this protocol, the term gender will be used to refer to the biological differences between males and females.

Multiple social variables inclusive of gender have shown to influence sex differentials in how diseases progress, access individuals have to health services, and health outcomes. Moreover, gender role expectations are generalizations or assumptions made on individuals' biological sex (9), which have adverse influence on the decision-making power of the individual, their autonomy, access to resources and health facilities as well as their power in society (14). For instance, women have been identified in some conditions as more vulnerable to poor health outcomes than men are as a result of gender-based inequities in access to income and other resources (3,14). These gender vulnerabilities often affect men and women in inverse ways with new literature conveying that COVID-19 had more adverse health effects for Chinese men in comparison to the women, evident by the 2.7% fatality rate in men,

compared to the 1.7% in women. In addition to the difference in fatality rate, European countries have reported more adverse outcomes in older men in comparison to older women (8).

The influence of gender has been emphasized in recent literature with its significance in specified health outcomes, differences in pharmacology and immunity playing an essential role in understanding sex and gender differences in health outcomes (10). These gender-based differences in immunity, specifically in COVID-19 infections have indicated increased contributions to the severity of the infection, and mortality in men, with an increase in the presentation of prolonged symptoms in women (15). Furthermore, in COVID-19 data reporting gender-disaggregated data has found to be lacking, as many investigators do not pre-plan gender-disaggregated data analysis within their studies (3). The need for reporting gender-disaggregated data is especially significant considering the knowledge gap regarding which gender is more adversely affected across countries (6).

Gender differentials in COVID-19 may be studied by numerous analytical methods, however the use of machine learning appears to be a more efficient approach. The use of machine learning techniques for the development of models for predicting COVID-19 trends has aided in early detection of infection and predicting pandemic peaks which have proven to reduce the burden on healthcare systems. Machine learning forms part of the advanced artificial intelligence (AI) concepts which provides complex and objective algorithmic techniques for multidimensional data analysis and assists with the diagnosis, detecting and monitoring of numerous diseases (17). These techniques can be classified in four different ways namely, supervised learning, unsupervised learning, weak supervision (semi-supervised) and reinforcement learning. The supervised learning approach consists of computer algorithms which are trained on labeled input data. The model takes raw labeled data to train and develop models necessary for predicting output values. The model can detect relationships between input data and output labels as well as yield results which are accurate from data that hasn't been seen before (16). The unsupervised approach in contrast is used for datasets which are not labeled, the algorithm is given unlabeled data and uses that data to detect patterns without any assistance. A

function is deduced to pull out hidden patterns from the unlabeled dataset (17). The weak supervision approach is most beneficial when there is a lack of quality data, or the dataset consists of imperfect labels. This approach lies in the center of supervised and unsupervised learning, it determines the correlations between data points and utilizes the labeled data to mark those data points (18).

The use of machine learning approaches for COVID-19 has proven to be highly beneficial. As seen in a pilot study conducted by the Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC) which provided multi-dimensional data on COVID-19 hospitalization and modelled fourteen features of individual-specific characteristics. Methods of supervised machine learning were used to develop a classification procedure. The study used a deep neural network (DNN) to train separate classes within the dataset, these classes focused on the severity of the illness and grouped data into severe illness and less severe illness respectively. The pilot study found that the responses to the spread of the COVID-19 disease are driven by numerous factors which include gender, age, and comorbidity (16).

Through the emerging use of machine learning researchers can gain an understanding of the role gender plays in health and specifically in COVID-19, it is essential to addressing knowledge gaps and identifying areas where interventions tackling gender-based vulnerabilities can be implemented for improved health outcomes. Addressing this gap in knowledge can be achieved by performing gender analysis, which requires the analysis of COVID-19 gender disaggregated data. This will play a paramount role in identifying any differences in COVID-19 health outcomes, inclusive of incidence rates, hospitalization, and mortality [14].

Problem statement:

In Gauteng province, the epicenter of the COVID-19 pandemic in South Africa, COVID-19 cases reported consistently higher in females than males (12). Suspected reasons could be poor health seeking behavior among men, meaning more women test for COVID-19 in comparison to men (13). Not much is known about gender differences in COVID-19 hospitalization and related outcomes, other

than women have had more cases of COVID-19 hospitalization in comparison to men, (16,19). Differences in COVID-19 hospitalization between men and women in Gauteng has been documented descriptively. However there has not been an in-depth analysis of correlations between gender and other variables which may influence the relationship between gender and COVID-19 hospitalization. For this reason, this study will use a weak supervision machine learning approach to provide gender-disaggregated data on COVID-19 hospitalization and in-hospital mortality for Gauteng province, determine if any observed gender difference is statistically significant, and analyze for hidden correlations to identify associated clinical and demographic factors.

Justification of the research

This research will form the basis for further research on whether gender-based vulnerabilities explain any observed gender differences. The implementation of weak supervision machine learning methods will be used as opposed to epidemiological methods as these machine learning methods enable learning from the data in a manner which is unbiased (17). Moreover, the methods which will be used are best suited to learning from data and understanding the gender differentials in COVID-19 hospitalizations and outcomes with minimal pre-existing knowledge (20). This study will therefore utilize aspects of machine learning to identify hidden associations which are necessary for informing health interventions and guiding policy makers.

Aim & objectives

Aim:

Using weak supervision machine learning techniques to identify gender differentials in COVID-19 hospitalizations, probabilities of hospitalization outcomes and hidden correlations with demographic and clinical factors that could account for the gender differences in COVID-19 hospitalizations in Gauteng.

Objectives are to:

1. Describe and compare the clinical and demographic characteristics of male and female COVID-19 patients admitted in Gauteng hospitals during March 2020 to July 2021.

2. Apply machine learning to determine if there is a statistically significant difference in COVID-19 hospitalization rates, disease severity, and mortality between males and females?
3. Apply machine learning to identify the hidden clinical and demographic variables that account for any observed gender differences in COVID-19 hospitalization and outcomes.

Methods

Research design and scope

A cross sectional study involving secondary data analysis will be best suited for analysing the data and providing support for further investigation on the topic. The proposed research will be limited to COVID-19 hospitalizations that occurred within the Gauteng province. The study population includes patients admitted with severe COVID-19 symptoms to private and public hospitals in the Gauteng province between the period of March 2020 and July 2021.

Data source

The National Institute for Communicable Diseases (NICD) has launched a sentinel hospital surveillance system named DATCOV, to monitor and describe trends of COVID-19 hospitalizations as well as the epidemiology of patients hospitalized within South Africa (13). The data were collected across all nine South African provinces from both private and public hospitals. Once informed, interested hospitals signed up and were authenticated (19). The hospitals reported on COVID-19 admissions and patient demographic data, clinical information (inclusive of any comorbidities patients had), clinical care, drugs received and outcomes. For this Masters research, secondary data analysis will be performed on this DATCOV national active surveillance system for COVID-19 admissions dataset, focussing only on the Gauteng dataset, including data for all three waves of the Gauteng pandemic (20). The NICD has shared the Gauteng dataset with the Gauteng Department of health (GDOH), from which personal identifiers have been removed and the dataset further shared with the University of Witwatersrand iThemba laboratory team (13). The analysis of the data for the sake of this Masters research will therefore be conducted with the support of the Wits iThemaba Lab team.

The data provided in the dataset has been anonymised and the dataset itself is password protected, restricting access to unauthorised individuals. Patient information contained in the dataset has been coded as a means of anonymizing the data, patient identifiers are also removed from the data. This study will therefore use aggregated patient information. Stored data is password protected and only shared with authorised individuals (19).

Variables

The dataset contains data about patients admitted with COVID-19 across hospitals in South Africa, the variables obtained have been grouped in categories namely, demographic (admission date, patient name, surname, ID number, date of birth, sex, race, pregnancy and occupation as a healthcare worker), comorbidities (hypertension, diabetes mellitus, cardiac disease, chronic pulmonary disease, asthma, chronic renal disease, HIV, TB and obesity), complications, treatment, setting of care (general ward, high care or intensive care unit) and outcomes (19). The variables which will be included in the analysis and assessed are provided in table 1. \

Table 4, variables to be included in analysis

Category	Variable
	Hospitalizations (No. and %)
	Rate of hospitalization (%)
	Admission date
Demographic variables	Age (median, IQR)
	Age categories
Setting of care upon admission	- ICU - High care - General ward
Care received on admission	- Oxygen not ventilated - Oxygen, ventilated - No oxygen (room air)
Comorbidities	- Number of comorbidities - Co-morbidities
Outcomes *(in-hospital mortality)*	- Length of stay before outcome - Discharged alive - Deceased

Overview of Proposed Solutions

As part of the proposed methodology, the Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC) will assist with aspects involving the machine learning modelling of the data. The consortium consists of an interdisciplinary team of statisticians, mathematicians and data scientists who are currently working on the designing COVID-19 vaccination strategies to minimise the number of COVID-19 cases which require hospital care (20). Weak supervision machine learning models will be used, this method utilizes machine learning algorithms in order to analyse groups of data which are imperfectly labelled. This method is highly beneficial as it uncovers hidden associations or patterns within the data, without the need for human interaction. Weak supervision machine learning is a great method for exploratory data analysis, which speaks to the main objectives of this proposed study. This method of data analysis includes minimizing human error, information bias and it is quick and easy to carry out in comparison to the time it would take to process the same number of features using epidemiological methods (21). Figure 1 provides a graphical explanation of the proposed solution.

Pre-processing

The dataset will be cleaned during the pre-processing stage, this will include the removal of null values, and incomplete or corrupted data. The pre-processing will remove any possible seasonality found in the dataset and identify the variables to include in the model.

Feature selection

Feature selection in machine learning refers to the process of selecting the most relevant features/ variables in our data to give to our model. The weak supervision feature selection method takes place through a density-based feature categorization in which two similar measures are used for continuous or discrete features (21). The categorization of patients will be done according to their risk and will be separated by gender.

Data splits

The dataset will be split to allow two subsets, one for training the model and the other for testing the performance of the model. This process will enable cross validation which seeks to achieve stable and confident estimates of how the model performs. For this study the data will be split with 70% of the data used for training, 20% used for testing and the remaining 10% will be used for validation.

Data analysis

Descriptive statistical analysis will be performed to address the first objective, the use of histograms will assist in visualizing the data and identifying patterns and correlations which are not apparent in the raw data. The variables which will be included are clinical and demographic namely: gender, age, comorbidities and the severity of illness. These variables will provide insight into the patients admitted by Gender. A threshold will be created for the comorbidities and the sum thereof will be used to categorize patients accordingly. For the second objective the hospitalization rates will be derived from the admission date, and a time series graph will be used to determine daily hospitalization rates. The measurement of the outcome of COVID-19 hospitalization will be dependent on the DNN output distributions, which will be divided into different classes based on the severity of their illness upon admission.

Weak supervision learning algorithm will be used to classify the dataset into two groups. The data associated with people who experienced more severe symptoms upon COVID-hospitalization will fall into the first group, this will include those admitted to intensive care units, high care u nits and those were succumbed to COVID-19 and/ related complications. The second group will consist of individuals who did not require intensive care due to their COVID-19 hospitalization. The variables which will be used to determine severity will be setting of care, care received on admission and outcome. The use of this technique will aid in discovering any similarities when the data is visualized. The test statistics used will be provided from the output of the machine learning techniques.

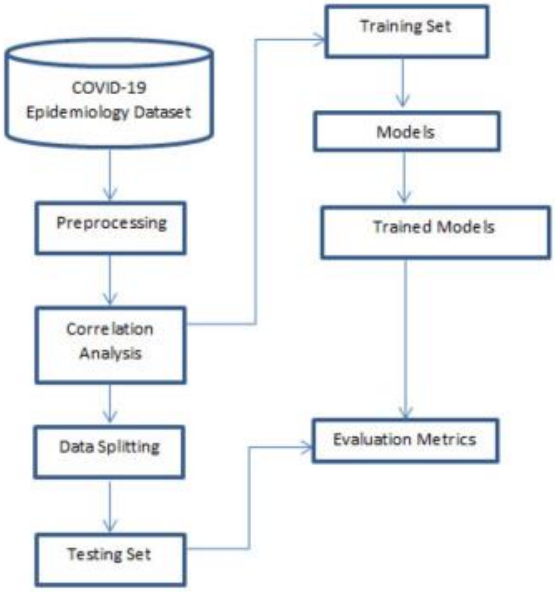

Figure 4, Overview of proposed solutions

Performance testing and optimization

The variables contained in the dataset will be selected as input variables to that will be used to prepare the Deep Neural Network (DNN). The DNN is a deep learning architectural method of machine learning which uses numerous data layers to extract higher-level features from raw data input. In this study the DNN will be used to classify individuals/ into two separate classes based on the input variables (setting of care, outcome etc.). The input variables will include comorbidities and gender. The test dataset (unseen) will then be used to allocate the individuals into the relevant classes.

Due to the multidimensional model which will be used, A receiver Operating Characteristic (ROC) curve will then be used to measure the performance of the classification by the DNN. From the ROC curve, the area under the curve can be calculated, this will provide the model's performance score. The higher the score the greater the performance of the DNN and the more accurate the results (20).

Performance evaluation

The evaluation of the performance of the model is essential for determining the goodness of the model. As part of the intrinsic metrics, the accuracy metric will be used to evaluate the performance of the model and determine how accurate the predicted data points are (21).

Environmental setup:

Anaconda will be used as the open-source Python distribution, where Python version 3.9 will be the programming language used for both the classification algorithm and data analysis. All programming will run on an Intel Core i3 processor 3.40 GHz CPU with 8 GB Random Access Memory (RAM), running the Windows 10 operating system. The association algorithm will be done using Python, containing a library called Keras that provides syntax for fulfilling different deep learning mechanisms. The following additional libraries will be imported into python:

Data cleaning and analysis

- Pandas
- Numpy

Data visualization

- Matplotlib
- Seaborn
- Roc_curve, roc_auc_score, auc

Building the model

- Sklearn
- Tensorflow
- Keras

Research limitations

Performing secondary analysis imposes limitations in the variables to be analyzed, in relation to the DATCOV dataset analysis is limited to variables present in the dataset. Other limitations include DATCOV being a sentinel surveillance system, therefore not all COVID-19 hospitalization have been included, the lack of reporting from some hospitals in the earlier stages of the pandemic were not represented within the dataset. Moreover, when looking at the outcomes for hospital admissions for COVID-19, the data only reported on hospital-based deaths which excluded any deaths which took place outside of the hospital.

Ethical considerations

This protocol will be submitted to the University of Witwatersrand Human Research Ethics Committee for ethics clearance. The de-identified DATCOV dataset which has been shard with the supervisors of this Masters research will only be used for the purpose of this research project and for publication at the University and peer review journals. Authorization to use the dataset will be obtained from the GDOH. The use of a dataset for which there was removal of personal identifiers, and password protection of the dataset will ensure that confidentiality will be maintained for all study participants.

Project timeline

Activity	Timeframe							
	2021					2022		
	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar
Submit protocol to Wits SPH Assessor Group	■	■						
Obtain permission for DATCOV dataset		■						
Ethics committee submission			■					
Data processing and analysis				■	■			
Report write-up					■	■		
Submission for examination							■	
Final submission								■

References

1. Witteman, Holly O., Jenna Haverfield, and Cara Tannenbaum. "COVID-19 gender policy changes support female scientists and improve research quality." *Proceedings of the National Academy of Sciences* 118.6 (2021).

2. Ekpanyaskul, Chatchai, and Chantana Padungtod. "Occupational Health Problems and Lifestyle Changes Among Novice Working-From-Home Workers Amid the COVID-19 Pandemic." Safety and health at work (2021)

3. Mittal S, Singh T. Gender-Based Violence During COVID-19 Pandemic: A Mini-Review. Frontiers in Global Women's Health. 2020;1.

4. Hawkes, Sarah, et al. "Recorded but not revealed: exploring the relationship between sex and gender, country income level, and COVID-19." *The Lancet Global Health* 9.6 (2021):

e751-e752. Quercioli, Cecilia, et al. "Gender differences in health expenditure determinants: Cecilia Quercioli." The European Journal of Public Health 26.suppl_1 (2016): ckw174-012.

5. Forrester JD, Forrester JA, Basimouneye JP, Tahir MZ, Trelles M, Kushner AL, Wren SM. Sex disparities among persons receiving operative care during armed conflicts. Surgery. 2017 Aug 1;162(2):366-76.

6. GDOH. Gauteng Hospitalization Dashboard.2021. Available online: https://www.covid19sa.org/gautenghospitalisation

7. Heidari S, Ahumada C, Kurbanova Z. Towards the real-time inclusion of sex- and age-disaggregated data in pandemic responses. BMJ Global Health. 2020;5(10):e003848.

8. Ya'qoub, Lina, Islam Y. Elgendy, and Carl J. Pepine. "Sex and Gender Differences in COVID-19: More to be learned!." American Heart Journal Plus: Cardiology Research and Practice (2021): 100011.

9. Gebhard, Catherine, et al. "Impact of sex and gender on COVID-19 outcomes in Europe." Biology of sex differences 11 (2020): 1-13.

10. Fatouros, Selina, and Teresa Capetola. "International journal of disaster risk reduction examining gendered expectations on women's vulnerability to natural hazards in low to middle income countries: A critical literature review." International Journal of Disaster Risk Reduction (2021): 102495.

11. Power, Kate. "The COVID-19 pandemic has increased the care burden of women and families." Sustainability: Science, Practice and Policy 16.1 (2020): 67-73.

12. Hankivsky, Olena, Kristen W. Springer, and Gemma Hunting. "Beyond sex and gender difference in funding and reporting of health research." Research integrity and peer review 3.1 (2018): 1-14.

13. Wan K, Tok P, Yoga Ratnam K, Aziz N, Isahak M, Ahmad Zaki R et al. Implementation of a COVID-19 surveillance programme for healthcare workers in a teaching hospital in an upper-middle-income country. PLOS ONE. 2021;16(4):e0249394.

14. Peckham H, de Gruijter N, Raine C, Radziszewska A, Ciurtin C, Wedderburn L et al. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. Nature Communications. 2020;11(1).

15. Lei Z, Lundberg S. Vulnerable Boys: Short-term and Long-term Gender Differences in the Impacts of Adolescent Disadvantage. Journal of economic behavior & organization. 2020 Oct 1;178:424-48.

16. Mellado B, Wu J, Kong J, Bragazzi N, Asgary A, Kawonga M et al. Leveraging Artificial Intelligence and Big Data to Optimize COVID-19 Clinical Public Health and Vaccination Roll-Out Strategies in Africa. International Journal of Environmental Research and Public Health. 2021;18(15):7890

17. Muhammad L, Algehyne E, Usman S, Ahmad A, Chakraborty C, Mohammed I. Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. SN Computer Science. 2020;2(1).

18. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics. 2020;ahead-of-print(ahead-of-print).

19. NATIONAL COVID-19 DAILY REPORT - NICD [Internet]. NICD. 2021. Available from: https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/surveillance-reports/national-covid-19-daily-report/

20. Rich-Edwards, Janet W., et al. "Sex and gender differences research design for basic, clinical, and population studies: essentials for investigators." *Endocrine reviews* 39.4 (2018): 424-439.

21. Katyara S, Staszewski L, Zbigniew L. Signal parameter estimation and classification using mixed supervised and unsupervised machine learning approaches. IEEE Access. 2020;:1-1.

22. Liu D, Zhou X. Covariate Adjustment in Estimating the Area Under ROC Curve with Partially Missing Gold Standard. Biometrics. 2013;69(1):91-100.

23. Valarmathi R, Sheela T. Heart disease prediction using hyper parameter optimization (HPO) tuning. Biomedical Signal Processing and Control. 2021;70:103033.