NEAR EAST UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

DEPARTMENT OF BIOSTATISTICS

ARTIFICIAL INTELLIGENCE (AI)-DRIVEN ENSEMBLE AND BOOSTING MODELS TO PREDICT COVID-19 MORTALITY IN EASTERN AFRICA

Ph.D. THESIS

Kedir Hussein ABEGAZ

Nicosia

September 2023

NEAR EAST UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

DEPARTMENT OF BIOSTATISTICS

ARTIFICIAL INTELLIGENCE (AI)-DRIVEN ENSEMBLE AND BOOSTING MODELS TO PREDICT COVID-19 MORTALITY IN EASTERN AFRICA

Ph.D. THESIS

Kedir Hussein ABEGAZ

Supervisor

Prof. Dr. İlker ETİKAN

Nicosia

September 2023

APPROVAL

We certify that we have read the thesis submitted by **Kedir ABEGAZ** titled "**Artificial Intelligence (AI)-driven ensemble and boosting models to predict COVID-19 mortality in eastern Africa**" and that, in our combined opinion, it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy in Biostatistics.

Examining Committee	**Name-Surname**	**Signature**
Head of the Committee (Supervisor)	Prof. Dr. İlker ETİKAN	
Committee Member	Prof. Dr. Selim Yavuz SANİSOĞLU	
Committee Member	Assoc. Prof. Dr. Özgür Tosun	
Committee Member	Prof. Dr. Beyza AKDAĞ	
Committee Member	Assoc.Prof. Dr. Uğur BİLGE	

16../11../2023

Prof. Dr. İlker Etikan

Approved by the Head of the Department — Head of Department

../11../2023

Prof. Dr. Kemal Hüsnü Can Başer

Approved by the Institute of Graduate Studies — Head of the Institute

DECLARATION

I hereby declare that all information, documents, analyses and results in this thesis have been collected and presented according to the academic rules and ethical guidelines of the Institute of Graduate Studies, Near East University. I also declare that as required by these rules and conduct, I have fully cited and referenced information and data that are not original to this study.

Kedir ABEGAZ

___/____/2023

ACKNOWLEDGMENTS

I would want to use this occasion to offer my heartfelt thanks to all of the people who helped me finish my PhD dissertation. Firstly, I am deeply indebted to my supervisor, **Prof. Dr İlker Etikan**, for his guidance, expertise, and unwavering support during this process. His insightful counsel, helpful critique, and encouragement have been pivotal in shaping the direction and calibre of my research. I also appreciate the expertise, inspiration, and ongoing support of the professors at the Department of Biostatistics. I also want to express my profound gratitude to my family, friends, and coworkers for their support, inspiration, and unshakable faith in me. Their compassion and emotional support helped me get through the difficulties encountered throughout this laborious procedure. Last but not least, I would like to express my gratitude to the participants who so kindly gave their time and wisdom, without whom this study would not have been possible. I appreciate the joint efforts of everyone involved because they were so crucial to the completion of my PhD dissertation.

Thanks to Allah	**Allah'a şükür**	**الحمد لله**	**ምስጋናየ ለአላህ ይሁን**

Kedir ABEGAZ

To all family members

ABSTRACT

Artificial Intelligence (AI)-driven ensemble and boosting models to predict COVID-19 mortality in eastern Africa."

Abegaz, Kedir Hussein

PhD, Department of Biostatistics

00/00/2023, 100 Pages

COVIID-19 severly affected Eastern Africa as of other parts of the world, which significantly disrupted social and economic activities in the region. This objective of this study was to predict mortality due to COVID-19 using artificial intelligence-driven ensemble and boosting models in Eastern Africa. The study used a two years daily data collected consecutively. At the preprocessing stage, the dataset was split into training and verification sets . In the modelling, sensitivity analysis, development of single black box AI-driven models, development of ensemble and boosting models, and comparison of ensemble models with single models were conducted. In the sensitivity analysis, four inputs were selected and used to run the single models, and accordingly, the coefficients of determination (DC) for ANFIS, FFNN, SVM, and MLR were, respectively, 0.9273, 0.8586, 0.8490, and 0.7956. Another four inputs were used to create the boosting method: AdaBoost, KNN, ANN-6, and SVM were shown to have determination coefficients of 0.9422, 0.8618, 0.8629, and 0.7171, respectively. The ANFIS ensemble and AdaBoost algorithms proved to be the most effective at enhancing the prediction performance of the single AI-driven models with non-linear ensemble techniques. This demonstrates the ability of ensemble and boosting models to predict COVID-19 mortality in Eastern Africa.

Keywords: Ensemble, Boosting, machine learning, covid-19, AdaBoost, ANFIS, FFNN

ÖZET

Doğu Afrika'da COVID-19 ölüm oranını tahmin etmek için Yapay Zeka (AI) güdümlü topluluk ve güçlendirme modelleri"

Abagaz, Kadir Hüseyin

Doktora, Biyoistatistik Anabilim Dalı

00/00/2023, 100 Sayfa

COVİD-19, dünyanın diğer bölgeleri gibi Doğu Afrika'yı da ciddi şekilde etkilemiş, bölgedeki sosyal ve ekonomik faaliyetlerde ciddi aksamalara yol açmıştır. Bu çalışmanın bu amacı, Doğu Afrika'da yapay zeka odaklı topluluk ve güçlendirme modellerini kullanarak COVID-19'a bağlı ölümleri tahmin etmekti. Çalışmada ardı ardına toplanan iki yıllık günlük veriler kullanıldı. Ön işleme aşamasında veri seti eğitim ve doğrulama setlerine bölündü. Modellemede duyarlılık analizi, tek kara kutulu yapay zeka destekli modellerin geliştirilmesi, topluluk ve güçlendirme modellerinin geliştirilmesi ve topluluk modellerinin tekli modellerle karşılaştırılması gerçekleştirildi. Duyarlılık analizinde dört giriş seçilmiş ve tekli modelleri çalıştırmak için kullanılmış ve buna göre ANFIS, FFNN, SVM ve MLR için belirleme katsayıları (DC) sırasıyla 0,9273, 0,8586, 0,8490 ve 0,7956 olmuştur. Güçlendirme yöntemini oluşturmak için dört girdi daha kullanıldı: AdaBoost, KNN, ANN-6 ve SVM'nin sırasıyla 0,9422, 0,8618, 0,8629 ve 0,7171 belirleme katsayılarına sahip olduğu gösterildi. ANFIS topluluğu ve AdaBoost algoritmalarının, tek yapay zeka destekli modellerin tahmin performansını doğrusal olmayan topluluk teknikleriyle geliştirmede en etkili algoritmalar olduğu kanıtlandı. Bu, birleştirme ve güçlendirme modellerinin Doğu Afrika'daki COVID-19 ölümlerini tahmin etme yeteneğini gösteriyor.

Anahtar Kelimeler: Ensemble, Boosting, makine öğrenimi, covid-19, AdaBoost, ANFIS, FFNN

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AdaBoost	Adaptive Boosting
AI	Artificial Intelligence
ANC	Antenatal care
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANFISE	Adaptive neuro-fuzzy inference system ensemble
ANN	Artificial Neural Network
AURROC	Area under the ROC curve
BFGS	Broyden, Fletcher, Goldfarb and Shanno
CNN	Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
CSSE	Center for Systems Science and Engineering
CT	Compound Tomography
CVD	Cardiovascular Disease
DC	Determination Coefficient
DL	Deep Learning
DM	Diabetes Mellitus
ECA	Economic Commission for Africa
ELM	Extreme Learning Machine
EMR	Electronic Medical Records
FFNN	Feedforward Neural Network
GDP	Gross Domestic Product
Grad-CAM	Gradient Weighted Class Activation Mapping
HOG	Histogram of Oriented Gradients
IMF	International Monetary Fund
JHU	John Hopkins University
KNN	K-Nearest Neighbors
MERS	Middle East Respiratory Syndrome
MFs	Membership Functions
ML	Machine Learning
MLP	Multilayer Perceptron

MLR	Multiple Linear Regression
NB	Naïve Bayes
NNE	Neural Network Ensemble
OWID	Our World in Data
RBF	Radial Basis Function
RF	Random forest
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SAE	Simple Average Ensemble
SARS	Severe Acute Respiratory Syndrome
SARS-CoV2	Severe Acute Respiratory Syndrome Coronavirus 2
SD	Standard Deviation
SDG	Sustainable Development Growth
SVM	Support Vector Machine
t-SNE	T-Distributed Stochastic Neighborhood Embedding
TT	Tetanus Toxoid
U5	under Five
UN	United Nations
WAE	Weighted Average Ensemble
WHO	World Health Organization

CHAPTER I

INTRODUCTION

1.1. Background of the study

"Artificial Intelligence could be the saviour of the COVID-19 pandemic in the coming year; we just need to prove it."

(The Lancet Digital Health, 2021)

Most pandemics in the 20th and 21st centuries were caused either by the coronavirus or the influenza virus. Among these pandemics, the viral disease pandemics that have occurred in the last twenty years are MERS, SARS, bird flu, H7N9, Ebola, H1N1, Nipah, and Zika (Arora et al., 2020). In December 2019, Wuhan, China, experienced the most recent coronavirus outbreak of this decade, called coronavirus disease 2019. This outbreak is one of the 21st-century pandemics and highly contagious infections caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) (Gao et al., 2021; Ko et al., 2020).

The World Health Organization (WHO), on 30 January 2020, declared this outbreak a "Public health emergency of international concern" and on 11 February 2020, named this outbreak 'COVID-19', used as a shorthand for coronavirus disease 2019. Finally, the WHO declared this outbreak a "Pandemic" on 11 March 2020 (Arora, Banerjee, & Narasu, 2020). The nature of this pandemic was different from earlier pandemic types and had a devastating effect on the world economy and led to a nearly complete cessation of social and economic activity worldwide (Abegaz & Etikan, 2022; Arora et al., 2020).

In terms of mortality, COVID-19 has caused more than 6.5 million deaths (6,559,902 as of 8 October 2022) globally, with a case fatality rate of 0.52% (WHO, 2020). This number proves that the pandemic is much different, in terms of global crises, compared with previous flu pandemics, such as the Spanish flu (in 1918), the Asian flu (1957–1958), the Hong Kong flu (1968–1970), and the swine flu (2009–2010). The nature of this pandemic made COVID-19 the first global public health issue that had a brutal impact on the global economy, which triggered a near-to-total shutdown of social and economic

activities. Finally, the pandemic has shrunk the global economy by nearly 3 percent, according to the prediction of the International Monetary Fund (IMF) (GAVI, 2022).

Remarkably, the crisis due to COVID-19 proves that our earth is unprepared for such a quickly spreading and rampant virus, resulting in a catastrophic pandemic. In addition to this, the big question, then, is, "When will things go back to normal, or whether we should prepare for new waves of coronavirus or not?" Though no one has a final answer to this question, through data analyses, we can understand how it happened and what the situation will look like in the future. The results of these analyses, including those using artificial intelligence (AI)-driven models and Boosting algorithms, will be actionable knowledge that can help us to manage a similar crisis in the future (Baik, Lee, Hong, & Park, 2022; Karaarslan & Aydın, 2021).

In this catastrophic era, AI is contributing to the development of many effective strategies that can control infection in real-time and easily track the rampant virus (Arora et al., 2020). It is also successfully used for the identification of disease, monitoring of cases and deaths, and prediction of future outbreaks and risks of mortality by analyzing the previous data of patients regarding the cases and deaths. In addition, AI can significantly boost the consistency of treatment and decision-making by developing important data-driven and machine-learning algorithms (Hu, Ge, Li, Jin, & Xiong, 2020; Vaishya, Javaid, Khan, & Haleem, 2020).

One of the most successfully recognized algorithms in the field of machine learning is the Adaptive Boosting (AdaBoost) algorithm, which was developed by Freund (1997). The AdaBoost algorithm, which maintains a collection of weights over training data and adjusts them after each weak learning cycle adaptively, creates a set of weak learners by assuming that a combination of weak learners can be "boosted" into an accurate strong learner (Freund & Schapire, 1997).

In contrast to conventional back-propagation neural networks or convolutional neural networks, recent examples of research have shown that AdaBoost-based machine learning could achieve high accuracy in modelling with multi-class imbalanced data (W. Sun & Gao, 2019; Taherkhani, Cosma, & McGinnity, 2020). AdaBoost has been used in

ensemble learning because of its superior classification and prediction performance, which includes image recognition, estimation of fruit biochemical parameters, and complex change prediction modelling (Fernandes et al., 2011; Liu, Tian, Li, & Zhang, 2015; J. Sun, Fujita, Chen, & Li, 2017; Zhao, Gong, Zhou, Huang, & Liu, 2016). In addition to the Boosting algorithm, the prediction capability of different non-linear AI-driven superior neural networks models like SOFNN-HPS and GK-ARFNN was found to be high in predicting wastewater treatment processes (Zhou, Li, Zhang, Xu, & Su, 2022; Zhou, Zhang, Duan, & Zhao, 2020).

Furthermore, web applications were developed by Chowdhury D. et al. in 2022. The application can detect whether a patient has COVID-19 or not after the image of the chest X-ray they uploaded to the web application (Chowdhury et al., 2022). Through AI-supported imaging technology, unenhanced chest computed tomography (CT) becomes applicable to the prediction of COVID-19. According to Schiaffino S. et al., in 2021, multilayer perceptron was the best-performing machine learning algorithm in predicting pulmonary parenchymal and vascular damage using unenhanced chest CT (Schiaffino et al., 2021).

1.2. Statement of the problem

The COVID-19 pandemic has had a significant effect on East Africa, with serious economic repercussions and endangering the region's growth and progress toward sustainable development. In this region, communicable diseases were the leading causes of mortality in the earlier stage of the COVID-19 pandemic, and among these diseases, perinatal, maternal, and malnutrition cases were responsible for almost half of the mortality in the region. In addition to this, East Africa is facing momentous health-related challenges due to preventable infectious diseases.

However, it is still facing a challenge due to the pandemic, and mortality remains at an alarming rate. This rate is likely to increase in the coming years because of COVID-19 and its consequences in the region, and Ethiopia is one of the significant contributors to the region’s mortality (UN. ECA, 2022). The resulting crisis and the pandemic itself threaten to reverse the development of some parts of the region that occurred within the last decade

and will hinder progress toward sustainable development growth (SDG) (UN. ECA, 2022; Web, 2022).

In the region including Ethiopia, many studies produced information on different healthcare issues such as antenatal care (ANC) utilization status of mothers (Abegaz & Habtewold, 2019), the postnatal care (PNC) visits of mothers (Sahle, 2016), access to tetanus toxoid (TT) immunization of mothers (Abegaz & Atomssa, 2017), predicting undernutrition status of U5 children (Markos, Doyore, Yifiru, & Haidar, 2014), predicting the CD4 count status of patients under ART (Mariam & Mariam, 2015), predicting the level of anaemia among women (Dejene, Abuhay, & Bogale, 2022), and predicting U5 mortality (Bitew, Nyarko, Potter, & Sparks, 2020) by using different machine-learning algorithms and AI-driven models. In addition to this, a few studies have applied them to the detection and classification of COVID-19 cases from X-ray images (Ayalew, Salau, Abeje, & Enyew, 2022; Erdaw & Tachbele, 2021).

Even though many studies (Dong et al., 2021; Gao et al., 2020; Guo & He, 2021; Kolozsvári et al., 2021; Ullah, Moon, Naeem, & Jabbar, 2022; Yaşar, Çolak, & Yoloğlu, 2021) have produced information regarding COVID-19 by using the concept of big data, machine learning and artificial intelligence, studies related to the prediction of mortality using different AI-driven models are rare globally and in the region. According to our search of various databases, no study has discussed the use of ensemble modelling and boosting algorithms to predict COVID-19 mortality in the region. This study had conducted in response to the stated problems.

1.3. Purpose of the study

The purpose of this study was to develop and evaluate AI-driven ensemble and boosting models specifically designed for predicting COVID-19 mortality in Eastern Africa. The study aims to achieve the following specific objectives:

- Evaluate and select suitable AI-driven models, including ensemble modelling and boosting algorithms, for COVID-19 mortality prediction

- Assess the performance and predictive accuracy of the developed AI-driven ensemble and boosting models
- Compare the performance of ensemble and boosting models against single and weak learner models commonly used in COVID-19 mortality prediction.
- Conduct a comparative analysis of the AI-driven ensemble and boosting models to identify the most effective and reliable model for COVID-19 mortality prediction in Eastern Africa.

1.4. Significance of the study

This study bears important significance for several stakeholders and offers the scholarly community the following contributions:

- **Addressing a Critical Need**: COVID-19 had serious effects on public health, economics, and development. The study tackles this critical and pressing need. The research's goal is to deliver precise and timely predictions of COVID-19 mortality in the area through the development of AI-driven ensemble and boosting models. The implementation of targeted interventions, resource allocation, and preventative actions to lessen the impact of the pandemic on public health may be made easier with the use of these data by healthcare authorities, policymakers, and organizations.
- **Filling a Research Gap:** There are few pieces of research available on mortality prediction using AI-driven models in Eastern Africa, despite the increased interest in AI and machine learning for COVID-19 analysis. The paper closes this research gap and provides insightful information on the use of AI-driven ensemble and boosting models, especially for the region's COVID-19 mortality prediction. The results will increase our understanding of this field and lay the groundwork for future investigation and development of AI-driven models in epidemiology and healthcare research.
- **Enhancing Prediction Accuracy**: The project intends to increase the precision of COVID-19 mortality estimates in Eastern Africa through the application of ensemble modelling and boosting methods. Ensemble models can integrate the

advantages of many AI-driven models, producing forecasts that are more solid and trustworthy. The research will shed light on the efficiency and superiority of boosting algorithms in capturing the intricate patterns and dynamics of COVID-19 mortality in the area by contrasting their performance against that of single and weak learner models.

- **Informing Policy and Decision-Making**: Accurate prediction of COVID-19 mortality can have major effects on public health initiatives, resource allocation, and policy development. Policymakers, healthcare workers, and other stakeholders involved in controlling the epidemic in Eastern Africa can benefit greatly from the study's results. The predictions made by AI-driven models can help with proactive planning, risk assessment, and targeted actions to avoid or reduce death rates, which will eventually save lives and lessen the pandemic's social and economic effects.
- **Technological Advancement**: The study advances AI-driven algorithms and models in the fields of epidemiology and healthcare. The work demonstrates the ability of AI to deliver accurate and reliable forecasts in complicated and quickly moving circumstances like the COVID-19 pandemic by utilizing cutting-edge methodologies like ensemble modelling and boosting algorithms. The findings of this study may encourage future investigation and use of AI-driven models in healthcare systems, enhancing such systems' capacity for illness monitoring, management, and response.

1.5. Limitations of the study

While conducting, several limitations should be acknowledged:

- **Data Availability and Quality:** The study's conclusions mainly rely on data on COVID-19 patients, mortality, demographics, comorbidities, and socioeconomic variables in Eastern Africa. The validity and generalizability of the findings may be harmed by limitations in the data collection, accuracy, completeness, and representativeness.
- **Generalizability:** Since the study focused on Eastern Africa, the conclusions won't be immediately relevant to other continents or nations. The dynamics of COVID-

19, such as mortality patterns, the incidence of comorbidities, and the state of the healthcare system, might fluctuate across different geographical regions, limiting the generalizability of the proposed models.

- **Model Performance:** Ensemble and boosting models driven by AI are subject to several restrictions in terms of their accuracy and predictive performance. Model overfitting, selection bias, feature engineering, and model assumptions can all affect the model's performance and bring uncertainty into mortality estimates.
- **Data Representation and Bias:** The dataset used to train the AI models may have inherent biases, such as an under or overrepresentation of particular areas or demographic groups. The models' capacity to correctly forecast COVID-19 mortality for certain subpopulations may be impacted by biases in data collecting, testing procedures, and reporting methods.
- **The dynamic nature of COVID-19:** The COVID-19 pandemic is a fast-changing scenario that is characterized by shifting epidemiological patterns, appearing variations, and shifting healthcare approaches. The results of the study could be a reflection of the circumstances and trends that prevailed at a particular time, but they might not be an accurate representation of prospective changes in death rates or disease dynamics over time.
- **Interpretability of AI Models:** AI-driven ensemble and boosting models frequently have complicated structures and are difficult to understand. It might be difficult to offer thorough justifications for mortality forecasts due to our limited understanding of the underlying systems and factors affecting the model's predictions.
- **External Factors:** Although the study focuses on the use of AI-driven models for mortality prediction, it does not take into account all of the interventions and external factors that may affect COVID-19 mortality. Beyond the purview of this study, factors like vaccination drives, public health initiatives, and healthcare system capacity may influence death rates.

For the study's findings and implications to be correctly interpreted, these limitations must be acknowledged. The validity and generalizability of the findings should be

discussed in full transparency, and researchers should make every effort to resolve these constraints as best they can.

CHAPTER II

REVIEW OF LITERATURE

In this chapter, both theoretical and empirical literature related to this study has been searched and presented. Even though a lot of research is conducted on COVID-19 datasets using different AI-driven models, Machine learning algorithms and data mining and knowledge discovery process, very little literature are available on AI-driven ensemble and boosting models. Therefore, this Literature review summarizes different works related to the COVID-19 pandemic using different machine-learning algorithms and their prediction performance.

2.1. Theoretical Framework

In this sub-section, the theoretical framework, the philosophy of the ensemble model and common types of ensemble modelling are presented.

2.1.1. Philosophy of Ensemble model

Several theoretical and empirical studies have demonstrated that the performance of ensemble and boosting models outperforms the performance of single and weak-performing models (Abdunabi, T.A., 2016). Ensemble systems work for three general reasons: statistical, computational, and representational (Thomas G.D., 2000) as follows:

- **Statistical algorithm:** The main purpose of this algorithm is to search the hypothesis's space to find the best hypothesis in the space.
- **Computational algorithm:** The justification of computational algorithms in applying ensemble systems include imperfect learning algorithms like too de32much data, small sample size, and data fusion. Some learning algorithms are only guaranteed to converge to local optima, such as the back-propagation method used to train neural networks (Kuncheva, 2014).
- **Representational**: The representational explanation is based on the fact that, in some situations, none of the space's hypotheses can adequately reflect the genuine hypothesis.

Although it is quite likely that the fusion of several models will perform better than a single model, the fusion of various models could just increase the system's complexity without improving performance. Therefore, for an ensemble to be successful, each base model must be unique.

2.1.2. Common Types of Ensemble Modelling

In this section, three common methods are reviewed, namely, bagging, boosting, and random subspace. More details on these methods can be found in the provided references.

a. Bagging

Bagging was introduced by Beriman, and it is sometimes known as "Bootstrap AGGregatING" (Leo B., 1996). By training the ensemble models using bootstrap replicates (sampling without replacement) of the training dataset, bagging aims to introduce model diversity at the data level. The outputs of the models are then averaged or blended using a majority vote for classification or, in the case of regression.

b. Boosting

Boosting is defined as the general problem of producing a very accurate prediction rule by combining rough and moderately inaccurate rules of thumb (Yoav F. et al., 1997). Due to their accuracy, applicability and robustness, boosting algorithms have been ranked among the top ensemble models. AdaBoost is one type of boosting algorithm, and it stands for "ADAptive BOOSTing". Resampling and reweighting are two types of implementing the AdaBoost algorithm (Yoav F. et al., 1996).

c. Random Subspace

The random space approach fixes the model at the feature (variable) level. Hence, models are trained on a randomly selected subset of features to construct the ensemble model (Tin K. H., 1998). Therefore, the final prediction from the ensemble model will be made by combining the outputs of models using a majority vote for classification purposes or averaging for regression. In addition to bagging, the random forest applies the random space method (Leo, B., 2001).

d. Stacking

The aim behind stacking is to integrate many weak learners by training a Meta model to provide predictions based on the various predictions that these weak models have returned. Stacking differs from bagging and boosting primarily in two ways. First, whereas bagging and boosting primarily take homogenous weak learners into account, stacking frequently take into account heterogeneous weak learners (various learning methods are merged). Second, while bagging and boosting combine weak learners using deterministic algorithms, stacking learns to combine the basic models using a Meta-model (Rocca J., 2019).

2.2. Related Literature

This part of the Literature review summarizes different works related to the COVID-19 pandemic using different ensemble machine learning algorithms and their prediction performance over single and weak performer models. However, a limited number of studies are discussed in this section due to the small number of literature available on ensemble modelling to predict COVID-19 mortality.

Kumar, K., investigated a case study on an ensemble model-based machine learning to predict the mortality risk of patients due to COVID-19. Three classifiers-NB, RF and SVM-were used in this study's analysis, and the ensemble techniques bagging and boosting were applied to increase the accuracy of the prediction. The findings show that adding the boosting method (AdaBoost) to an ensemble of weak classifiers significantly improves predictions for extremely heterogeneous datasets. Therefore, the boosting algorithms increase the prediction of the weak performer model called the SVM algorithm by 9.86 percent (Kumar K., 2021).

Ko, H. and Chung H. et al. developed an AI model to predict mortality at the early stage of hospital admission due to blood test data. They developed a COVID-19 AI model called EDRnet (Ensemble learning model based on deep neural network and random forest) to predict in-hospital mortality. The developed EDRnet model provided high sensitivity

(100 percent) and high specificity (91 percent), and an accuracy of 92 percent (Ko, H., Chung, H., 2020).

Baik, S. M. and Lee, M. et al. performed a study to analyze the electronic medical record (EMR) data and laboratory results of hospitalized COVID-19 patients in Korea using DL and ML to develop an optimized ensemble model to predict mortality. Three machine learning models (RF, SVM and XGBoost) and one DL model (Multilayer perceptron) were developed. Accordingly, they found that the ensemble model was the best-performing model with AUC=0.8811 and an accuracy of 0.85, and hence, this result demonstrated that the ensemble model had the highest ability of prediction in classifying COVID-19 mortality using EMR in Korea (Baik, S. M. and Lee, M. et al., 2022).

Ullah F. et al. developed an ensemble model to identify COVID-19 disease based on two datasets containing 1,646 and 2,481 CT scan images. In addition to the ensemble model, they conducted SVM, Decision tree Gaussian Naïve Bayes, K-nearest neighbour, logistic regression and Random forest machine learning models. They applied Gradient-weighted Class Activation Mapping (Grad-CAM) and t-distributed Stochastic Neighbor Embedding (t-SNE) to interpret the overall performance of the proposed models. Accordingly, the proposed ensemble approach outperformed other existing models with 98.5 percent accuracy and 99 percent precision (Ullah, F., 2022).

Ayalew et al. developed the hybrid model of detection and classification for quick diagnostics of COVID-19 disease using X-ray images collected from the university of Gondar database. In this study, one hybrid model called DCCNet and two single models, a convolutional neural network (CNN) and histogram of oriented gradients (HOG), were proposed for quick diagnosis of COVID-19 disease using X-ray images of patients. The experiment was conducted in a framework known as Keras (Transflow as the backend) using Python. According to the result, the hybrid model achieved 99.67 percent accuracy in detecting and classifying the disease, and this implies that the hybrid model (DCCNet) single models by 6.7 percent (Ayalew, A. et al., 2022).

Lou et al. developed an ensemble model using the computational method to predict mortality due to COVID-19 from 4,711 reported cases confirmed as SARS-CoV-2

infected. Their computational model (AURROC) was developed, combining machine learning and genetic algorithms using 10 features. According to their finding, the ensemble model performed better than others, with a ROC value of 0.7802. This shows the robustness of the ensemble model developed by combining machine learning and genetic algorithms (Lou L. et al., 2023).

Cui, S. et al. collected data from 79 countries to develop a model that predicts the development trend of mortality due to COVID-19 in these countries. Therefore, they developed three single models; Multiple linear regression (MLR), support vector machine (SVM), and extreme learning machine (ELM), and a two-layer nested, heterogeneous ensemble model combining three single models was developed. According to the results of the study, the proposed ensemble model shows better prediction ability than state-of-the-art machine learning methods (MLR, SVM, and ELM) (Cui, S. et al., 2021).

CHAPTER III

METHODOLOGY

3.1. Study Area

The study area of this research was Eastern Africa, which is a part of sub-Saharan Africa comprising two traditionally recognized regions: East Africa (Kenya, Tanzania, and Uganda); and the Horn of Africa (Somalia. Djibouti, Eritrea, and Ethiopia). This region is the most populous sub-region of Africa, representing nearly 5.6% of the world's population. It stretches from Mozambique in the south to Eritrea in the north. There are eighteen countries and two independencies in the region, but nearly a quarter of the region's people are living in one country, Ethiopia (USAID, 2022; World Atlas, 2022). Ethiopia is the second and first most populous country in Africa and Eastern Africa, respectively (Statista, 2022). Currently, the main focus under the health infrastructure development of the region includes the standardization and expansion of hospitals among states prioritizing the prevention and containment of COVID-19 (ITA, 2023).

3.2. Data Source and Attribute Selection

This study used COVID-19-related data collected daily over 24 months, from April 2020 to April 2022, in the region. The data were from the "Our World in Data (OWID)" team and the COVID-19 data warehouse at John Hopkins University (JHU), collected by the Center for Systems Science and Engineering (CSSE), which is open to users. Data were retrieved from: https://github.com/owid/COVID-19-data/tree/master/public/data (accessed on 25 June 2022). The OWID, in its statement under the license section, explains that "All visualizations, data, and code produced by 'Our World in Data' is completely open access under the Creative Commons BY license. You have the permission to use, distribute, and reproduce these in any medium, provided the source and authors are credited" (Hasell et al., 2020; Mathieu et al., 2021).

3.3. Data Preprocessing and Analyses

In the 'Our World in Data (OWID) COVID-19 database, many variables were available, but ten variables (for ensemble modelling) and seven variables (for boosting algorithms) were selected because of the availability and completeness of their data and their theoretical relationship with mortality due to COVID-19. For ensemble modelling, the datasets were retrieved for each country in the East Africa region independently and calculated the average values each day to represent the region with a single variable for mortality due to COVID-19 and other input variables. However, to boost the algorithm, Ethiopia's COVID-19-related datasets were directly retrieved from the source for pre-processing purposes. It is known that the daily collected data is non-linear by its nature. Hence, the first step before modelling was to check normality and normalize each non-normal variable in the datasets.

The second activity was to select the dominant input variables through a non-linear sensitivity analysis called the coefficient of determination (DC). This analysis was conducted using an artificial neural network (ANN), applying a bivariate analysis (one target and one input variable) to predict the estimated values and calculate the coefficient of determination to verify the correlation of each input variable with the target variable. The datasets from these countries were classified into two separate datasets: the training dataset (70 percent of the data) and the testing dataset (30 percent of the data) for the development of AI-driven ensemble models and boosting algorithms.

For all the developed AI-driven ensemble models and boosting algorithms, the target variable was the daily number of new deaths due to COVID-19 in the region, and the input variables were the new daily number of cases in the region, the positive rate, the number of people vaccinated, hospital beds/1000 patients, and so on. In Table **1**, the list of all the variables used in this study and their explanations are presented.

In the data processing step, the data normalization was calculated using Microsoft Excel and **MATLAB (Version 20)** was applied to conduct the sensitivity analysis, the single black-box AI-driven models, and AI-driven ensemble models. However, for

Boosting algorithms, **Orange data mining (Version 3.33)** was applied to model all weak performer models and the boosting algorithm.

Table **1:**

The target variable and input variables for this study.

Variables	The Description of Variables
New deaths	New deaths attributed to COVID-19
New cases	New confirmed cases of COVID-19.
Positive rate	The share of COVID-19 tests that are positive
People vaccinated	Total number of people who received at least one vaccine dose
Stringency index	A composite metric based on 9 reaction indicators, such as school closures, workplace closures, and travel prohibitions, rescaled to a score between 0 and 100 (100 is the strict response)
GDP per capita/USD	Gross domestic product at purchasing power parity
Number of smokers	Share of male and female smokers
Prevalence of DM	Prevalence of people with diabetes aged 20 to 79
Hospitals beds/1000	Hospital beds per 1000 people
Population density	Number of people divided by land area, measured in square kilometres

3.4. Proposed Methodology

In this section, the proposed methodology for both ensemble modelling and boosting algorithms is described clearly and presented as follows.

3.4.1. Proposed method for ensemble modelling

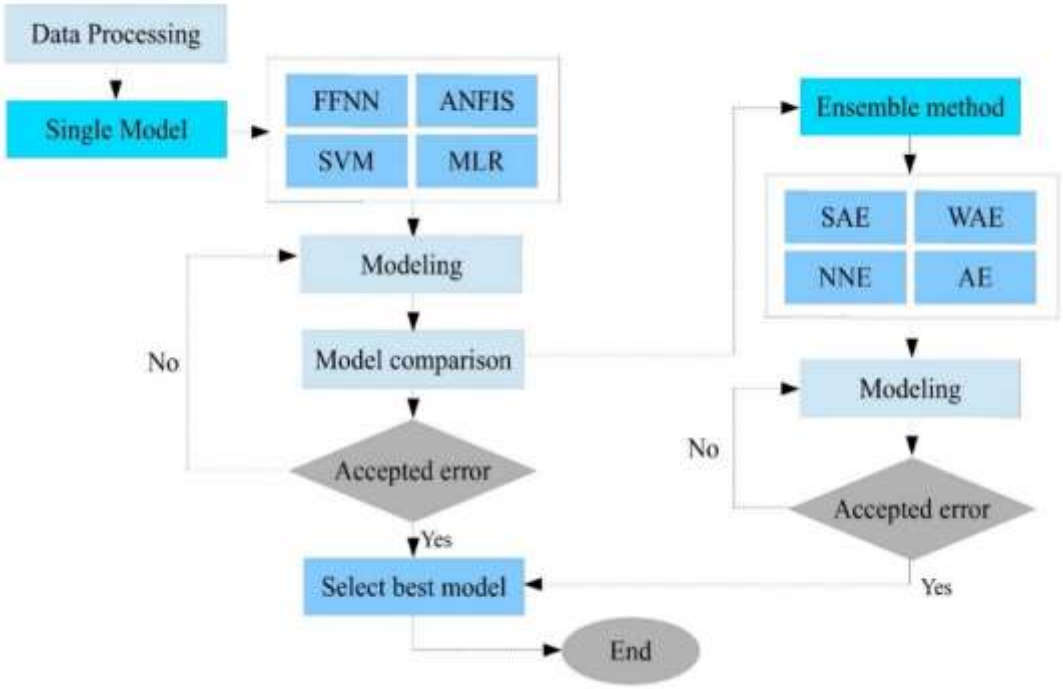
In ensemble modelling, the study modelled three AI-driven models, including an adaptive neuro-fuzzy inference system (ANFIS), feedforward neural network (FFNN), support vector machine (SVM), and one conventional data-driven model, multiple linear

regression (MLR), to predict mortality due to COVID-19 in East Africa. In addition, the dataset was classified into a training dataset and a test dataset after normalizing inputs.

Figure **1** shows the architecture of the overall ensemble modelling of the study. This architecture comprises three stages, which were conducted to carry out the given study. Firstly, the sensitivity analysis was conducted to rank and select the most influential input variables for the modelling. In the second stage, four AI-driven black-box models (ANFIS, SVM, FFNN, and MLR) were applied independently to predict COVID-19 mortality. Thirdly, as a final stage, four ensemble approaches, namely the ANFIS ensemble (ANFISE), neural network ensemble (NNE), weighted average ensemble (WAE), and simple average ensemble (SAE), were constructed. In the ensemble model stage, the estimated output of every single model was used as an input for the AI-driven ensemble modelling. Then, the predicted mortality based on the ensemble model was compared with the predicted results from each of the black-box models in the second stage.

Figure **1:**

Schematic diagram of the AI-driven models' development



The detailed description of all single AI-driven models and the ensemble models are presented in the following sub-sections.

a. The Feedforward Neural Network (FFNN)

The artificial neural network (ANN) is one of the most significant AI-driven models because it can build links between the target and input variables by training the neural network without having comprehensive information on the entire data set (Tanty & Desmukh, 2015). This model is a self-learning simulation function that demonstrates the capacity to model and forecast complicated processes.

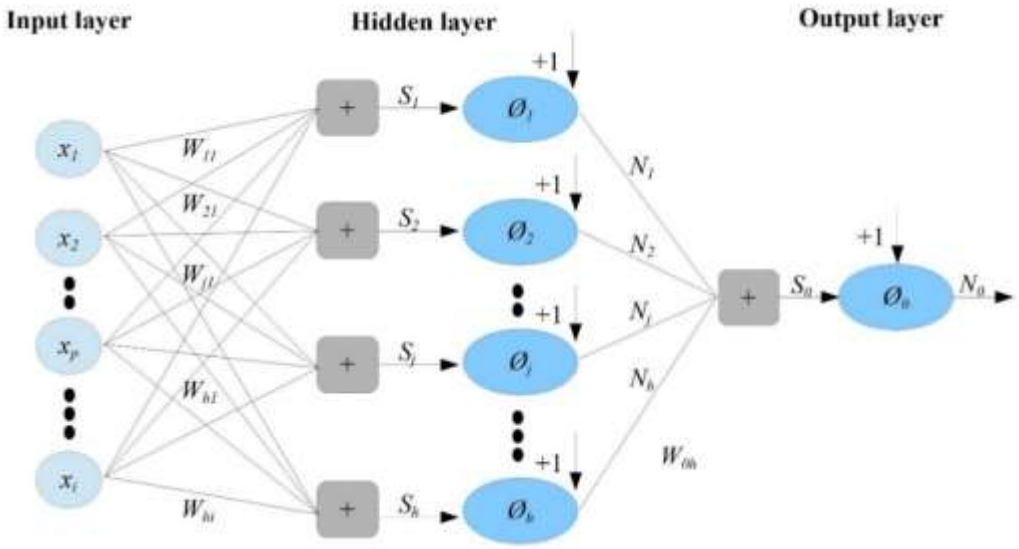
This capability makes ANN a more practical and efficient model in different domains of science, such as biomedical technology, engineering, agriculture, and business (Nourani, Gökçekuş, & Umar, 2020).

Because of its simplicity and favoured ability to react to various challenges without considering the past information regarding the process, this study used the feedforward neural network (FFNN), employing propagation algorithms.

FFNN is formed of linked pieces called 'nodes' that have unit properties of information, such as learning, non-linearity, noise tolerance, generalization capability, and so on, and it has three layers (see Figure **2**), including the input, the hidden, and the output layers. As a result, the input variables provided to the input layers 'neurons' are transmitted forwards, and the activation function, a non-linear function, is employed to construct the output vector.

Figure **2:**

The structure of the FFNN.



A multi-layer perceptron (MLP) model with a single hidden layer was computed in this FFNN model. The formal definition of this model is as follows: the function 'f' on the fixed-size input 'x', such as f(x) ≈ y for training pairs of (x, y). Alternatively, recurrent neural networks learn sequential data, computing the output 'Ø' on the variable-length input $X_k = \{x_1 \ldots x_k\} \approx y_k$ for training pairs of (X_n, Y_n) for all $1 \leq k \leq n$.

In the definition of FFNN with the 'm' layer (or 'm-2′ hidden layers) prototype, the output perceptron has an activation function $Ø_o$, and the hidden layer perceptron has activation functions Ø. Every perceptron in layer l_i is connected to every perceptron in layer l_{i-1}. The layers are fully connected, and there is no connection between the perceptrons in the same layer. According to Brilliant (Brilliant.org, 2022), it is computed using the following three steps:

First, initialize the input layer l_0 and set the values of the outputs $Ø_i^0$ for nodes in the input layer l_0 concerning their associated inputs in the vector $\vec{x} = \{x_1 \ldots x_n\}$, i.e., $Ø_i^0 = x_i$

Second, compute the sum of the products and each output of the hidden layer in the order from l_1 to l_{m-1} for 'k', progressing from 1 to m-1

- compute $h_i^k = w_i^{\rightarrow k} \varnothing^{\rightarrow k-1} + b_i^k = b_i^k + \sum_{j=1}^{r_{k-1}} w_{ji}^k \varnothing_j^{k-1}$, for i = 1…$r_k$

- compute $\varnothing_i^k = g(h_i^k)$, for i = 1…$r_k$

Third, compute the output y for the output layer l_m

- Compute $h_1^m = w_1^{\rightarrow m} \varnothing^{\rightarrow m-1} + b_1^m = b_1^m + \sum_{j=1}^{r_{m-1}} w_{j1}^k \varnothing_j^{k-1}$

- Compute $\varnothing = \varnothing_1^m = g_\varnothing(h_1^m)$, where the MLP uses the denotations below.

The w_{ij}^k is the weight for perceptron j in the layer l_k for the incoming node i, b_i^k is the bias for the perceptron i in layer l_k, h_i^k is a product of some plus bias for perception i in layer l_k, $\varnothing_i^k$ is the output for node i in layer l_k, r_k is several nodes in layer lk, $w_i^{\rightarrow k}$ is the weight vector for perceptron i in layer l_k, and $\varnothing^{\rightarrow k}$ is the output vector for layer l_k.

b. The Adaptive Neuro-Fuzzy Inference System (ANFIS)

The ANFIS is developed by combining the ability to learn the neural network and its advantage of a rule-based fuzzy inference system, which enables it to integrate a past observation into the process of classification (J.-S. Jang, 1993; J.-S. R. Jang, Sun, & Mizutani, 1997). This combination makes ANFIS a good model for overcoming the limitations of individual modelling. Jang JS, in 1997, described the 'defuzzifier', 'fuzzifier', and 'fuzzy' databases as the three parts of a fuzzy system (J.-S. R. Jang et al., 1997). Even though they are different from each other, the well-known and commonly used fuzzy inference systems are Mamdani's system (Mamdani & Assilian, 1975), Tsukamoto's system (Tsukamoto, 1979), and Sugeno's system (Takagi & Sugeno, 1985).

The ANFIS architecture contains five layers: Layer 1 is the input layer, Layer 2 is the input membership function (MFs), Layer 3 is the association rules, Layer 4 is the output membership function, and Layer 5 is the model output (see Figure 3). After the construction of the fuzzy system, it specifies the relationship between the fuzzy variables using the **'if-then'** fuzzy rules. The first order of Sugeno's system has the following rules,

considering that the fuzzy inference system contains a single output (f) and two inputs (x and y):

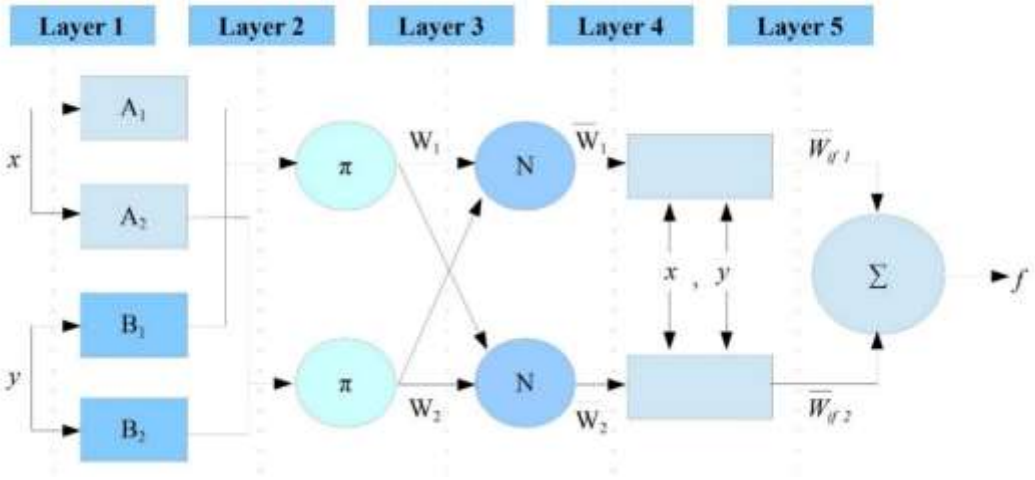
$$\text{Rule } (1)\text{: if } \mu(x) \text{ is } A_1 \text{and } \mu(y) \text{ is } B_1, \text{ then } f_1 = p_1 x + q_1 y + r_1$$

$$\text{Rule } (2)\text{: if } \mu(x) \text{ is } A_2 \text{and } \mu(y) \text{ is } B_2, \text{ then } f_2 = p_2 x + q_2 y + r_2$$

Where 'A' and 'B' are membership functions, and 'p', 'q', and 'r' are parameters for the outlet functions. Considering the stated parameters, the structure of the ANFIS with five layers is presented in Figure 3 and explained as follows:

Figure **3:**

The structure of the ANFIS



Layer 1: every i's node is an adaptive node that has the following node function on this layer:

$$Q_i^1 = \mu_{Ai}(x) \text{ for } i = 1, 2 \text{ or } Q_i^1 = \mu_{Bi}(x) \text{ for } i = 3, 4$$

Where Q_i^1 is for input 'x' or 'y', that is the membership grade. Here, the Gaussian membership function was selected because it has the lowest error of prediction.

Layer 2: In this layer, the 'T-Norm' operator connects every rule using the 'AND' operator between the inputs and is presented as:

$$Q_i^{\,2} = w_i = \mu_{Ai}(x).\,\mu_{Bi}(y) \text{ for } i = 1,2$$

Layer 3: In this layer, the output is the 'Normalized firing strength', and the labelled norm for every neuron is as follows:

$$Q_i^3 = \bar{w} = \frac{w_i}{w_1 + w_2} \quad i = 1,2$$

Layer 4: In this layer, every i's node is an adaptive node and executes the consequence of the rules by considering p_1, q_1, and r_1 as irregular parameters, as follows: $Q_i^{\,4} = \bar{w}(p_i x + q_i y + r_i) = \bar{w} f_i$

Layer 5: In this layer, the overall output is calculated by summing all the incoming signals as follows:

$$Q_i^{\,5} = \bar{w}(p_i x + q_i y + r_i) = \sum w_i f_i = \frac{\sum w_i f_i}{\sum w_i} \ldots\ldots (1)$$

c. The Support Vector Machine (SVM)

According to A. M. Kalteh, the support vector machine (SVM), may be utilized for both prediction and classification purposes (Kalteh, 2013; Schiaffino et al., 2021). The type of regression in the SVM model is known as support vector regression. It is used to define regression using SVM and structural risk reduction. Figure 4 depicts the framework of the SVM regression technique that can simulate non-linear situations in the real world. The estimation obtained using this regression is important for estimating a function of the given dataset:

$$(x_i d_i)_i^n$$

where x_i is the input vectors, d_i is the actual values, and n is the total number of the dataset. Hence, SVM has the following regression function:

$$y = f(x) = \omega\phi(xi) + b$$

where 'φ' is a non-linear mapping function, and 'ω 'and 'b' are parameters of the function of the regression that can be determined by assigning positive values for the slack parameters of ξ and ξ* and the minimization of the objective function, considering 'c' as the regularized constant and $\frac{1}{2}\| w \|^2$ as the weight vector norm, as shown below.

The Minimization:

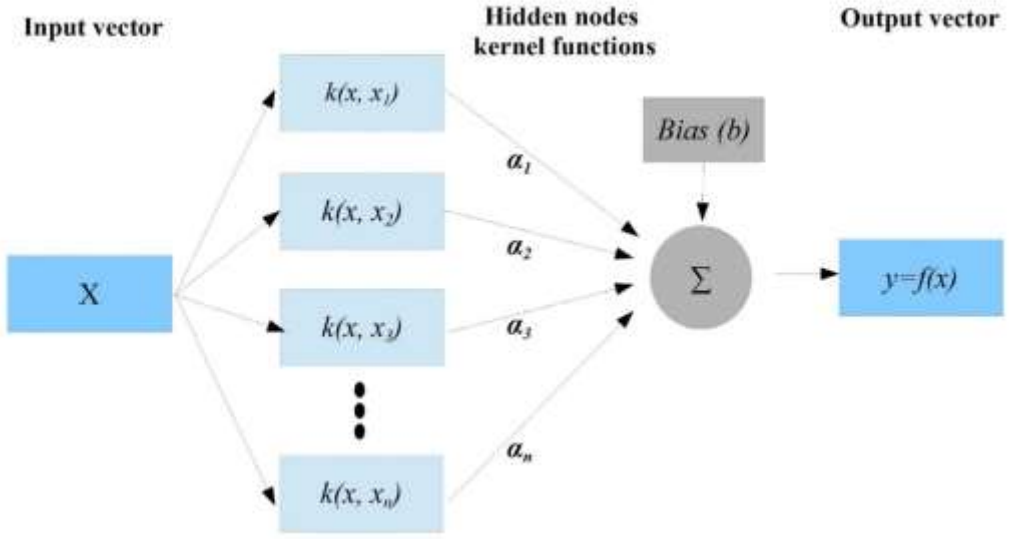
$$\frac{1}{2}\| w \|^2 + c\left(\sum_{i}^{n} \left(\xi_i + \xi_i * \right) \right)$$

This is subject to

$$\begin{cases} w_i\phi(x_i) + b_i - d_i \le \varepsilon + \xi^* \\ d_i - w_i\phi(x_i) + b_i \le \varepsilon + \xi^*, i = 1, 2, ...n \\ \xi\xi^* \end{cases}$$

Figure **4:**

The structure of SVM



The optimization problem stated above could be improved to obtain a dual quadratic problem of the optimization, defining the lag-range multipliers α_i and α_i*. In addition, the vector 'w' can be computed by identifying the optimization solution problem as follows:

$$w^* = \sum_{i}^{n} (\alpha_i - \alpha_i^*)\varphi(xi)$$

Hence, the SVM regression function is changed to:

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)k(x_i, x_j) + b$$

Where b is the bias term and k (x_i, x_j) is the kernel function that can be expressed as:

$k(x_1, x_2) = \exp(-\gamma \| x_1 - x_2 \|^2)$, where γ is the parameter of the kernel function.

d. The Multiple Linear Regression (MLR)

Multiple linear regression is commonly used as a statistical modelling technique to observe the linear relationships between numerically measured variables. It is a form of linear regression used to examine the linear relationship between a single target variable and several input variables. In this technique, the target variable (Y) is supposed to be affected by the input variables (X_i), and the estimated model can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

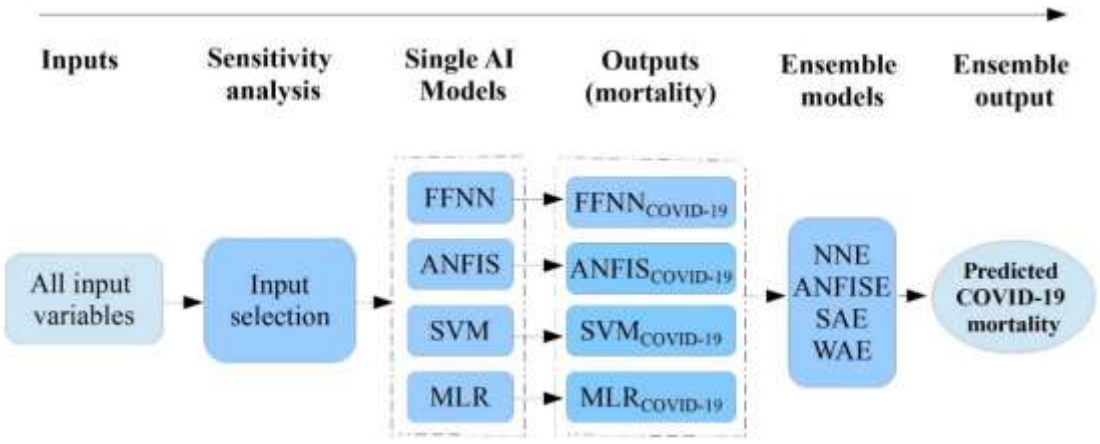
Where 'y' is the target variable, β_0 is the regression constant, β_i are coefficients of the input variables, and x_i are input variables.

3.4.2. Ensemble Modelling

In the AI industry, ensemble modelling is computed by combining the estimated predictions of multiple single AI-driven models. This combination can advance the final model's performance, and it can provide better predictions than the individually constructed models (Sharghi, Nourani, & Behfar, 2018).

Figure **5:**

Diagram of the ensemble process



This study used two linear ensemble techniques to increase the performance of AI-driven single models: the weighted average ensemble (WAE) and simple average ensemble (SAE), and another two non-linear ensemble techniques: ANFIS ensemble (ANFISE) and neural network ensemble (NNE), were applied (see Figure 5).

These ensemble techniques have been applied in various studies for purposes such as the clustering and classifications of medical data, web ranking, economic forecasting, etc. (Abba et al., 2020; Ajami, Duan, Gao, & Sorooshian, 2006; Edeh et al., 2022; Kazienko, Lughofer, & Trawiński, 2013; Y. Wang et al., 2021; Wu et al., 2021). Considering this situation, this study also applied the ensemble technique to predict COVID-19 mortality in East Africa.

a. The Linear Ensemble Approaches

In this approach, the simple average (SA) and weighted average (WA) ensemble techniques were applied. In the simple average technique, the arithmetic average of the output, the estimated COVID-19 mortality of every single AI-driven model is taken as the final predicted mortality in the region. Meanwhile, in the weighted average technique, the prediction is computed by assigning weights to each output relative to its importance.

The formula for the simple average: $\text{COVID}-19 = \frac{1}{N}\sum_{i=1}^{N}\text{COVID}-19_i$, where COVID-19 is the output of the SA ensemble model, and COVID-19_i is the output of i^{th} single AI-driven model. The formula for weighted average: $\text{COVID}-19 = \sum_{i=1}^{N} w_i \text{COVID}-19_i$, where w_i is a weight for the output of i^{th} method and is computed using the performance measure called the determination coefficient (DC) and can be calculated with $w_i = \frac{DC_i}{\sum_{i=1}^{n} DC_i}$, where DC_i is the coefficient of determination for the i^{th} model.

b. The Non-Linear Ensemble Approaches

In this approach, the non-linear averaging was computed by training the single AI-driven non-linear models (FFNN and ANFIS) using the COVID-19 mortality values predicted by the single AI-driven models, and the neural network ensemble (NNE) and ANFIS ensemble (ANFISE) were applied. In NNE, the non-linear averaging was performed by training different FFNNs by feeding the output of an AI-driven single model as an input. Then, the maximum epoch number and neurons of the hidden layer were determined by trial and error. Meanwhile, in ANFISE, the predicted COVID-19 mortality based on the AI-driven single model is fed to ANFIS for training using a different number of epochs and membership functions (MFs).

c. Normalization and Evaluation of Models

Both the target and input variables should be standardized before training the model at an early stage to reduce dimensions and guarantee that all the variables receive equal attention (Nourani, Elkiran, & Abba, 2018). The following normalization formula was performed on the dataset to establish the standardized values within the range of 0–1:

$$\mathrm{COVID-19}_{n} = \frac{(\mathrm{COVID-19})_{i} - (\mathrm{COVID-19})_{min}}{(\mathrm{COVID-19})_{max} - (\mathrm{COVID-19})_{min}},\ i = 1 \ldots n$$

COVID-19n, COVID-19i, $\mathrm{COVID\text{-}19}_{min}$, and $\mathrm{COVID\text{-}19}_{max}$ represent the normalized, actual, minimum, and maximum COVID-19 mortality values, respectively. Even though the best model for the validation and training steps is determined by trial and error (Dawson, Abrahart, & See, 2007), the root mean square error (RMSE) and determination coefficients (DC) were used to measure the performance and efficiency of the developed models. The DC values range from −1 to 1, and a model value approaching 1 yields better results. In addition, the model with the lowest RMSE is considered to be the best model. The formula of RMSE is presented as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((\mathrm{COVID-19})_{obsi} - (\mathrm{COVID-19})_{prei}\right)^{2}} \text{ and}$$

$$\frac{\sum_{i=1}^{N}\left((\text{COVID}-19)_{\text{obsi}}-(\text{COVID}-19)_{\text{prei}}\right)^2}{\sum_{i=1}^{N}\left((\text{COVID}-19)_{\text{obsi}}-(\overline{\text{COVID}}-19)_{\text{obs}}\right)^2}$$

where COVID-19$_{\text{obsi}}$, COVID-19$_{\text{prei}}$, $(\overline{COVID}-19)_{obs}$, and N are the observed COVID-19 mortality value, predicted COVID-19 mortality value, average of the observed COVID-19 mortality values, and number of observations, respectively.

3.4.3. The proposed method for boosting algorithms

The overall proposed methodology of boosting algorithm is presented in Figure **6** as a model development workflow of **Orange (Data mining).** It includes the pre-processing of datasets, normalizing of variables, sensitivity analysis, dividing data into training and testing datasets, model development, and prediction process on both training and testing datasets.

Once the data pre-processing was completed, three weak learner AI-driven models: k-nearest neighbours (KNN), the artificial neural network (ANN-6), and support vector machine (SVM), and one boosting ensemble model (Adaptive Boosting) were developed to predict mortality due to COVID-19 in Ethiopia's dataset.

Finally, the prediction performance of three AI-driven models was compared with the boosting model based on their result of the coefficient of determination (DC) and the root mean square error (RMSE).

The AI-driven models used in the boosting algorithm to predict COVID-19 mortality were the k-nearest neighbours (KNN), the artificial neural network (ANN-6), and the support vector machine (SVM). The boosting model used was the adaptive boosting (AdaBoost) AI-driven model. While developing these models, parameters selected, for each model, after trial and error are presented in Table **2** below.

Figure **6:**

Orange (Data mining) workflow of the proposed methodology

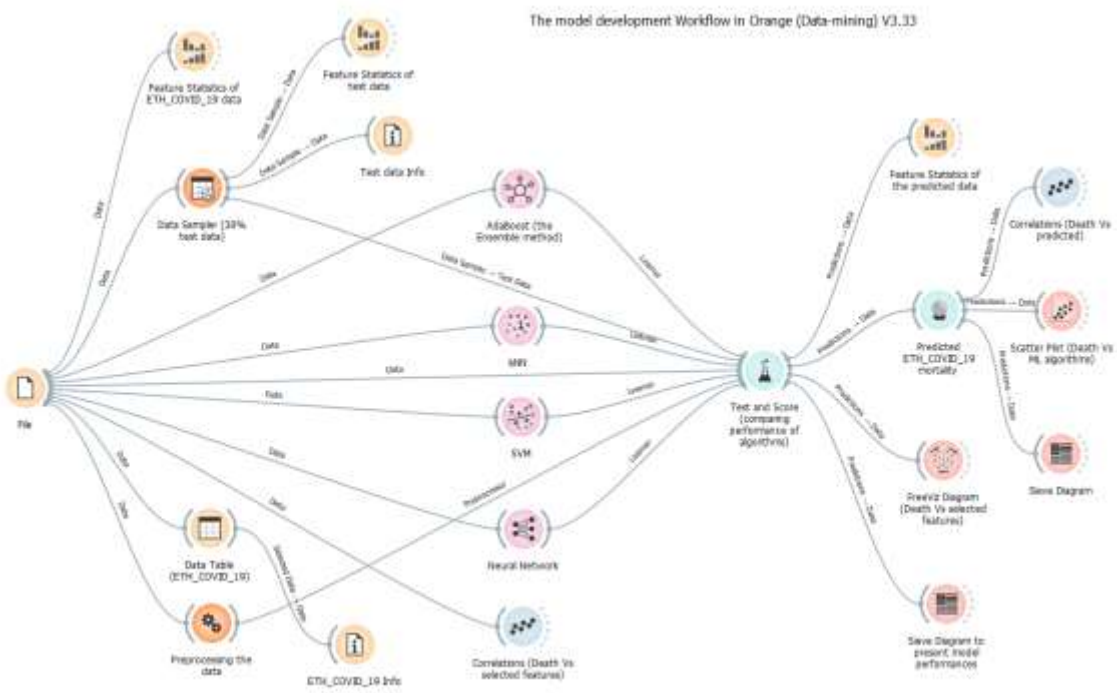


Table **2:**

Model parameters used to build AI-driven model criteria

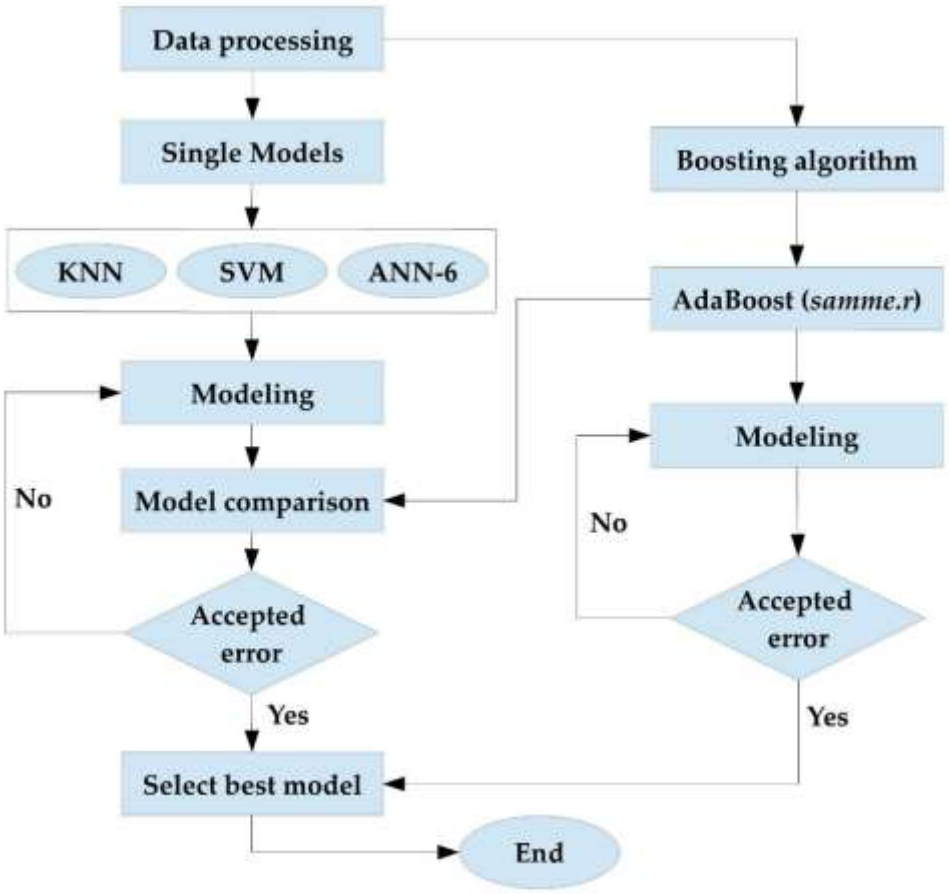
Models	Model parameters
AdaBoost	Base estimator: tree, Number of estimators: 4, Algorithm: Samme.r, and Loss (regression): Square
KNN	Number of neighbours: 2, Metric: Manhattan, and Weight: Uniform
SVM	SVM type: SVM, C=1.0, ε=0.10000000000000003, Kernel: RBF, exp(-auto|x-y|²), Numerical tolerance: 0.001, and Iteration limit: 300
ANN-6	Hidden layers: 200, Activation: tanh, Solver: L-BFGS-B, Alpha: 1, Max iterations: 500, and Replicable training: True

Figure **7** shows the flowchart for the model development process. This flowchart demonstrates the construction of models. The dataset was first preprocessed in the model

development process, after which individual models (KNN, SVM, and ANN-6) were developed separately and evaluated against the boosting algorithm (AdaBoost).

Figure **7:**

Flowchart of model developments.



a. Adaptive Boosting Regression (AdaBoost Regression)

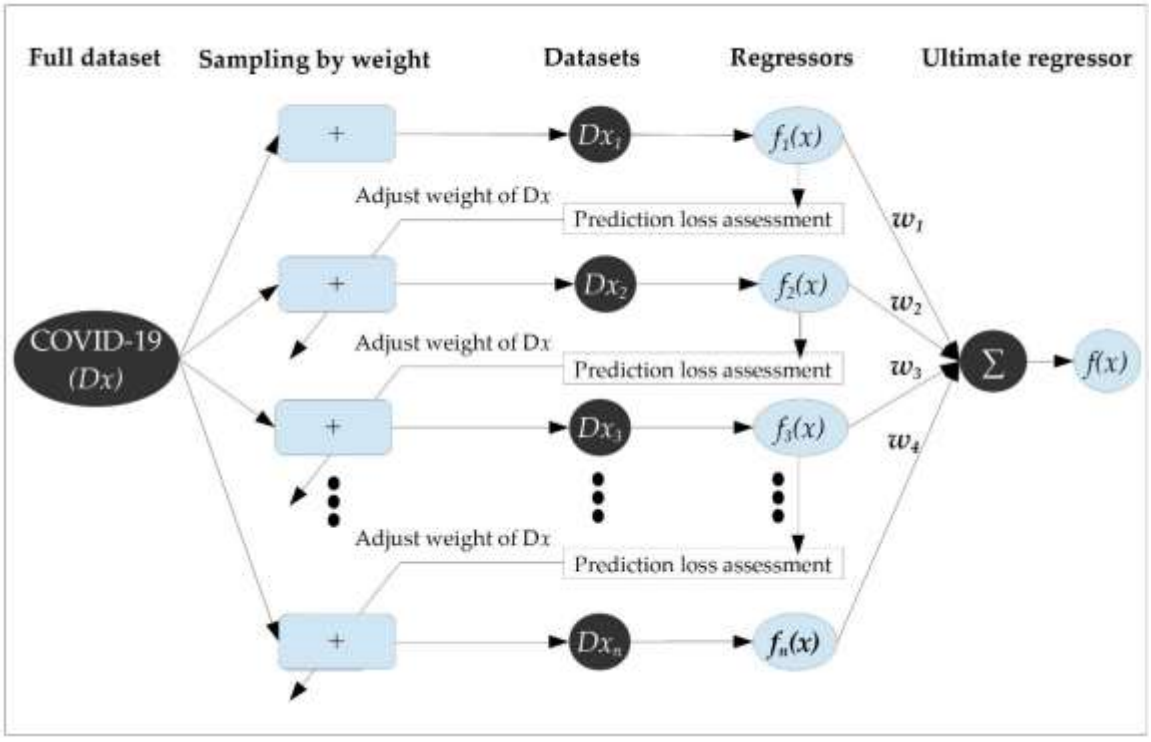
AdaBoost-based regression is a type of boosting AI-driven model that can apply a powerful machine learning algorithm for the regressing of target and feature variables (Freund, Schapire, & Abe, 1999; Rojas, 2009; Schapire, 2013). The purpose of applying a boosting regression was to obtain the best prediction from the ensemble of multiple weak predictors. The schematic presentation of the AdaBoost Model is presented in Figure **8**.

As we can observe from Figure **8**, the model processes the input COVID-19 dataset and denote this dataset as D_x. Initially, each dataset of D_x was assigned an equal weight, and this weight determined the chance of being sampled. Due to this weight, the model selected the training dataset (Dx1) from dataset D_x with replacement sampling, and hence, to train the regressor $f_1(x)$, the training dataset was used. As we can see from the schematic presentation, the prediction weight assessment was applied to assess the trained regressor 1 [$f_1(x)$] and calculate the weight 'w_1' for the regressor.

This assessment is to adjust the weight for the main dataset D_x. In the weighting process, the larger the prediction error, the larger the weight for that specific trained dataset. Finally, the model parameter used in this study to build the AdaBoost model after a lot of trial and error was (Base estimator: tree, Number of estimators: 4, Algorithm: Samme.r, and Loss (regression): Square).

Figure **8:**

Schematic diagram of AdaBoost regression



As we can see from the schematic presentation, the prediction weight assessment was applied to assess the trained regressor 1 [$f_1(x)$] and calculate the weight 'w_1' for the regressor. This assessment is to adjust the weight for the main dataset D_x. In the weighting process, the larger the prediction error, the larger the weight for that specific trained dataset. Finally, the model parameter used in this study to build the AdaBoost model after a lot of trial and error was (Base estimator: tree, Number of estimators: 4, Algorithm: Samme.r, and Loss (regression): Square). (See Table 2)

Min H and Luo X. (2016) have summarized the overall procedure of AdaBoost in eight steps, and it is presented as follows (Min & Luo, 2016):

Step 1: The dataset D_x with training samples can be represented as: $\{(x_j, y_j)\}_{j=1}^{M}$

Step 2: Assign a distribution with equal weight; it is stated as $\{p_{ij} = \frac{1}{L} \mid i = 1,2,...K; j = 1,2,...M\}$ for each training sample starts from i=1 and starts the loop

Step 3: In the i^{th} iteration, the sample training data (M) $\{(x_j, y_j)\}_{j=1}^{M}$ will be replaced with p_{ij} and use the sampled data to train a regressor $g_i(x; \beta_i)$

Step 4: Calculate the prediction loss $L_j = L[y_{j,} g_i(x_j; \beta_i)]$ for each member of D_x, where $L_j \in [0,1]$. Also, calculate the weighted average of the loss $\bar{L}$.

$$L_j = \frac{L[y_{j,} g_i(x_j; \beta_i)]}{D}, D = \sup\{L_j\}, j = 1,2,...$$

$$\bar{L} = \sum_{j=1}^{M} p_{ij} L_j$$

Step 5: The weight of the regressor $g_i(x; \beta_i)$ will be calculated, and it can be presented by the following formula:

$$w_i = \frac{\bar{L}}{1-\bar{L}}$$

Step 6: If 'i', in step 5, equals the maximum number of iterations K, it will stop the loop and move to step 8.

Step 7: Updating the distribution weight of the dataset D_x by making i=i+1 in equation L_j at Step 4 and move to the loop:

$$p_{ij} = \frac{p_{ij} w_i^{1-\bar{L}}}{Z_i}$$

Where Z_i is a selected normalized factor, and hence, P_{ij} will be a random distribution.

Step 8: The obtained K regressors will be incorporated into a single regressor respective to their weight $\{w_i\}_{i=1}^{K}$, and it will have a formula:

$$g(x;\beta_{weight}) = \sum_{i=1}^{K} w_i g_i(x,\beta_i)$$

b. K-nearest neighbours regression (KNN regression)

KNN regression is one of the best-known and simplest non-parametric regression types and it does not explicitly assume the parametric form of the target variable (Gareth, Daniela, Trevor, & Robert, 2013). Given a prediction point of X_0 and the value for K, the KNN regression will first identify the K training observations, which are closest to X_0, represented by N_0. The KNN then estimates the target variable Y using the average of all the training responses in N_0. The small number of K provides the most flexible fit that has a low bias but high variance, and hence, the optimum value for K will depend on the bias-variance tradeoff.

It can be presented the prediction formula of KNN as follows: $Y = \frac{1}{K} \sum_{x_i \in N_0} y_i$

In this modelling, the model parameters for the KNN were decided after a lot of trial and error in the model development process, and hence, the parameters that make KNN predict better than other parameters were (Number of neighbours: 2, Metric: Manhattan, and Weight: Uniform) (See Table 2).

c. The artificial neural network (ANN-6)

Because it can establish a connection between feature variables and the target variable by training neural networks without detailed knowledge of the dataset, the 'ANN-6' is a class of AI-driven models and is regarded as the most important model (Tanty & Desmukh, 2015). In a variety of scientific fields, including biomedicine, technology, agriculture, and business, ANN is more effective and useful (Nourani et al., 2020).

This is because of its self-learning simulation function, which shows how ANNs can predict and model complex processes like the daily number of COVID-19 mortality. The ANN-6 with a forward propagation algorithm was chosen to predict COVID-19 mortality. The model parameters of the ANN-6 were Hidden layers: 200, Activation: tanh, Solver: L-BFGS-B, Alpha: 1, Max iterations: 500, and replicable training: True) **(**See Table **2)**.

The ANN-6 with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm and with three layers (input layer, hidden layer and output layer) was selected after repeated trial and error, assuming different parameters with optimum prediction capability. In addition to the trial and error, the BFGS has a proven performance even for non-smooth optimization (BogoToBogo., 2023), like the daily mortality of COVID-19.

d. The support vector machine (SVM)

SVM was applied to predict using the regression known as support vector regression with final parameters of the SVM model after a trial and error was conducted (SVM type: SVM, C=1.0, ε=0.10000000000000003, Kernel: RBF, exp (-auto|x-y|²), Numerical tolerance: 0.001, and Iteration limit: 300) (See Table **2)**.

d. Data normalization and model performance evaluation

Before modelling the boosting model, the standardization of the target variable and input variables was conducted to normalize the data into the standardized value. This standardization assures reducing dimensions among variables and having equal attention in the modelling process (Nourani et al., 2018).

The coefficients of determination (DC) and root mean square error (RMSE) were computed in the performance evaluation of models. Based on the determined values of RMSE and DC, the top-performing model was chosen. Therefore, a model with a DC value close to 1 and the lowest RMSE was deemed to be the best-performing AI-driven model.

CHAPTER IV

FINDINGS AND EXPERIMENTS

This study conducted ensemble modelling and boosting algorithms to predict COVID-19 in eastern Africa. Therefore, the result of this study is presented in different sub-sections as follows: For ensemble modelling, three AI-driven and one classical model were developed, namely ANFIS, SVM, FFNN, and MLR, respectively. In addition, for boosting the algorithm, Three AI-driven models (KNN, ANN-6, and SVM) and one boosting model (AdaBoost) were modelled.

All of these models were trained on 70 percent of the eastern Africa COVID-19 dataset and tested on 30 percent of this dataset. As a result, this section included reports on feature statistics, sensitivity analysis, development of AI-driven black box and weak-performing models, and comparison of these models with the ensemble model and boosting algorithms.

4.1. Feature statistics description

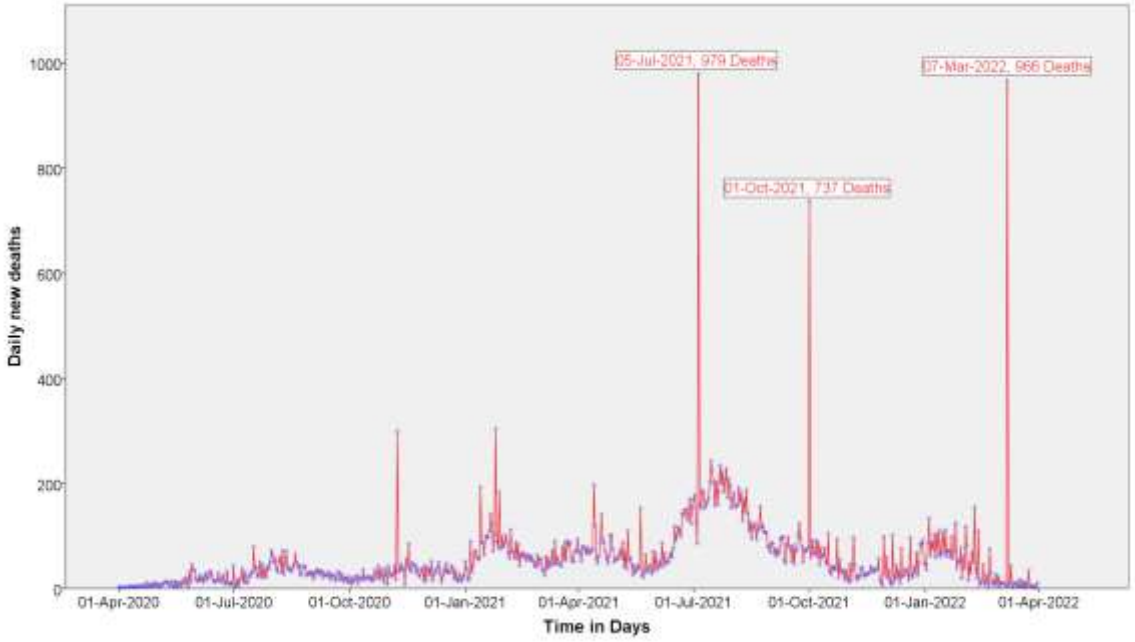
In this part of the result, the feature statistics of both ensemble modelling and boosting algorithms are presented subsequently.

4.1.1. Feature Statistics for ensemble modelling

A time series graph was used to present the data on the average daily mortality due to COVID-19 in East Africa, and it depicts in Figure **9**. In this graph, the three largest numbers of mortality cases observed per single day were 979, 966, and 737 on 5 July 2021, 7 March 2022, and 1 October 2021, respectively. However, more than 200 deaths per day were registered successively from July to August 2021, and we can conclude that this period was the peak time of mortality due to COVID-19 in the region.

Figure **9**:

Time series plot for the average daily mortality due to COVID-19 in the region.



The descriptive statistics in Table **3** present the average, maximum, and minimum values for the target and all input variables of the training (70 percent of the dataset) and verification datasets (30 percent of the dataset) from 01 April 2020 to 01 April 2022. The average mortality due to COVID-19 was (61.03 ± 69.1) for the training dataset, and it was (46.16 ± 83.59) for the verification dataset. The average number of new daily cases was (2783.5 ± 2423) for the training and (5724.66 ± 6522.36) for the verification data, while the rates of confirmed positive cases per day were 0.041 ± 0.022 and 0.05 ± 0.052 for the training and verification datasets, respectively.

The daily vaccine coverage and hospital beds per 1000 people, which are also presented in the table, showed that the average daily vaccine coverage was (26234.2 ± 47498.4) and (220514 ± 332466) for the testing and verification data, respectively. The hospital beds/1000 people were (28.25 ± 3.50) and (20.4 ± 0.5623) for the training and verification datasets, respectively.

Table **3**:

Descriptive statistics of the ensemble modelling on the COVID-19 dataset.

Variables	Training Data (n = 584)			Verification Data (n = 146)		
	Min	**Mean ± SD**	**Max**	**Min**	**Mean + SD**	**Max**
New deaths	0	61.03 ± 69.1	979	1	46.16 ± 83.59	966
New cases	11	2783.5 ± 2423	27596	95	5724.66 ± 6522.3	34125
positive rate	0.004	0.041 ± 0.022	0.102	0.0	0.05 ± 0.052	0.065
Newly vaccinated	0	26234.2 ± 47498.4	276532	2915	220514 ± 332466	1877713
Number of CVDs	4655.5	4822.25 ± 17.263	5231.50	4252	4656 ± 0.5268	4986
Stringency index	40.14	51.71 ± 8.80	76.50	29	40.80 ± 2.192	44
GDP per capita/USD	76254.4	76321.52 ± 2.35	76985.2	77956	76254 ± 2.589	78962
Number of smokers	354.2	365.5 ± 56.32	420.5	332.1	354.2 ± 9.536	386.5
Prevalence of DM	6.61	6.71 ± 0.23	6.98	6.51	6.61 ± 0.2652	7.02
Hospitals beds/1000	20.04	28.25 ± 3.50	35.23	18.1	20.4 ± 0.5623	22.6
Population density	2697.26	2725.25 ± 5.62	2756.85	2568.2	2697.25 ± 0.2562	2893.2

4.1.2. Feature statistics for boosting algorithms

The minimum, mean, maximum and standard deviation (SD) values of the target and feature variables are presented in Table **4** below for both training and testing datasets. In Ethiopia, the average number (mean ± SD) of daily mortality due to COVID-19 between 01 April 2020 and 01 April 2022 was (9.13±8.21) for the training dataset and (13.27±12.78) for the testing dataset. The average number of daily cases was (604.08±539.79) for the

training dataset and (756.02±1063.06) for the testing dataset. In addition to daily deaths and daily cases, the average bed capacity per/1000, the daily mask use (measured from 1), and the pneumonia status were (0.17±0.02), (0.42±0.16) and (0.96±0.09), respectively, for the testing dataset.

Table **4:**

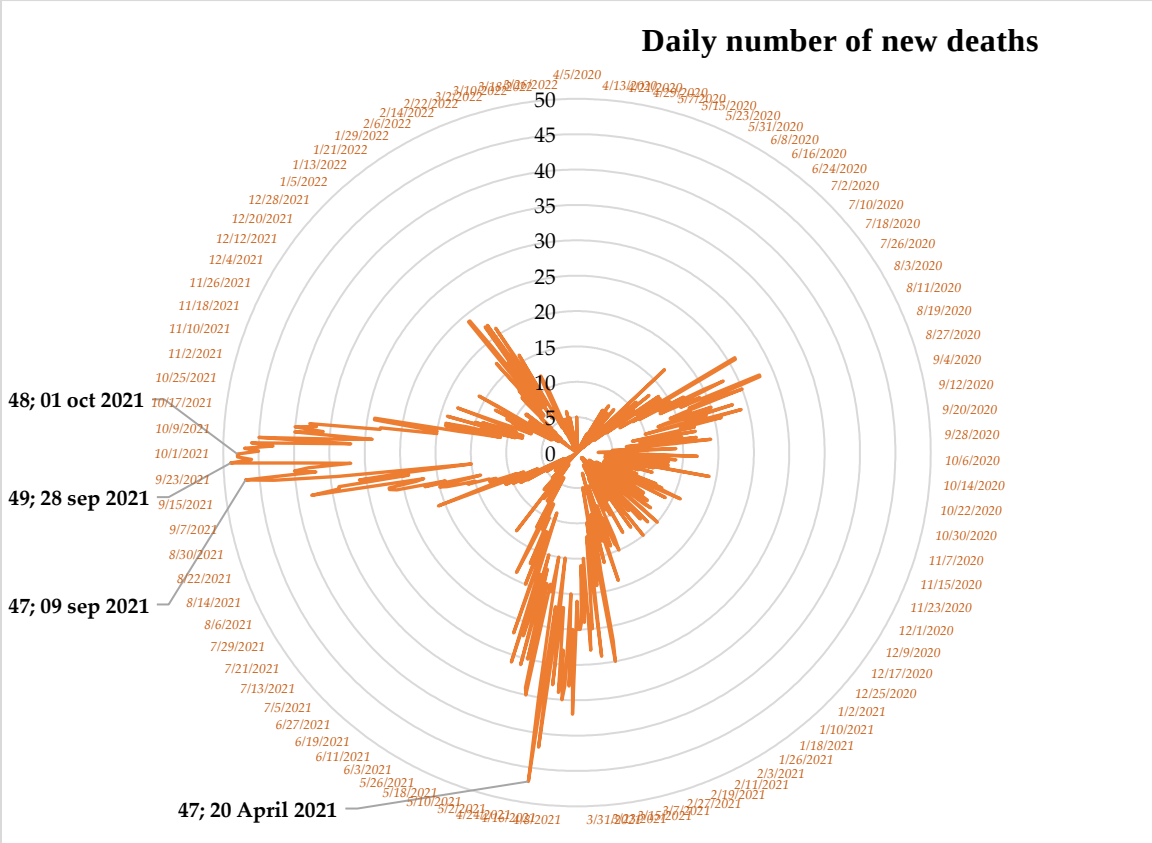
Descriptive statistics of the boosting model on the COVID-19 dataset

Variables	**Training Dataset**			**Testing Dataset**		
	Mean±SD	**Min**	**Max**	**Mean±SD**	**Min**	**Max**
New deaths	9.129±8.209	0	47	13.2667±12.777	0	49
New cases	604.08±539.8	0	2372	756.019±1063.1	7	5185
Bed capacity	0.1647±0.024	0.1245	0.186	0.1729±0.0213	0.135	0.1741
Mask use	0.4279±0.163	0.0000	0.669	0.4163±0.1641	0.000	0.8679
Pneumonia_st	0.9615±0.096	0.8213	1.093	0.9629±0.096	0.813	1.1294

The radar chart describes the daily number of new deaths in Figure **10**. In this chart, the recorded four largest numbers of daily mortality due to COVID-19 in Ethiopia were 49, 48, 47 and 47 deaths on 28 September 2021, 01 October 2021, 09 September 2021, and 20 April 2021, respectively. In addition to the largest number of daily deaths, 38 and more daily deaths were registered in the country from 13 September 2021 to 14 October 2021. Therefore, we can summarize that the peak time of COVID-19 mortality in Ethiopia was from 13 September 2021 to 14 October 2021.

Figure **10**:

Radar chart for daily COVID-19 mortality



4.2. Sensitivity Analysis

In this section, the sensitivity analysis of both ensemble modelling and boosting algorithms are presented subsequently.

4.2.1. Sensitivity analysis for ensemble modelling

The careful selection of the most relevant factors to consider as input variables and the correct adjustment of connecting parameters (such as the hidden neurons, number of iterations, and transfer functions) for any AI-driven modelling are crucial steps required to obtain the optimum prediction level. The number of new cases, rate of positive cases, newly vaccinated individuals, number of cardiovascular diseases (CVD), stringency index, GDP per capita in USD, number of smokers, the prevalence of diabetes mellitus (DM), Hospital

beds/1000, and population density were selected from the dataset for the sensitivity analysis based on their relationship with mortality and their completeness of data. Previously, linear sensitivity analytical approaches have been used to select dominant input variables. However, due to the complex non-linear nature of the COVID-19 data, it was conducted a sensitivity analysis with a non-linear nature. Hence, the non-linear FFNN was conducted to select the dominant input variables for the modelling of different AI-driven single and ensemble models to predict COVID-19 in eastern Africa.

The sensitivity analysis in this study to conduct the ensemble modelling is presented in Table **5**. Accordingly, the four best dominant input variables, with the four top highest Coefficient of determination values, selected were the positive rate (first-rank), hospital beds/1000 (second-rank), new cases (third-rank), and the number of vaccinated individuals (fourth-rank) based on their chronological order. However, the inputs with the lowest coefficient of determination (less than 0.5) were removed in the modelling process.

Table **5:**

Sensitivity analysis used to rank the inputs for ensemble modelling.

Inputs	DC	Rank
Positive rate	0.9178	1st
Hospital beds/1000	0.8962	2nd
New cases	0.8617	3rd
People vaccinated	0.8113	4th
Number of smokers	0.2505	5th
GDP per capita/USD	0.2220	6th
Number of CVDs	0.2013	7th
Population density	0.1902	8th
Prevalence of DM	0.0663	9th
Stringency Index	0.0524	10th

4.2.2. Sensitivity analysis for boosting algorithm

To obtain the optimum level of prediction of AI-driven models, the most important step is to carefully select the most relevant feature variables and adjust model parameters for every model. In the sensitivity analysis for boosting models, seven variables were included. These were 'mask_use', 'all_bed_capacity', 'new_cases', 'pneumonia_st', 'icu_bed_capacity', 'hosp_admission', and 'daily_infection'. Since the daily recorded data related to COVID-19 had a non-parametric nature, the neural network sensitivity analysis (the FFNN) was conducted to choose the dominant feature variables and is presented in Table **6**.

As we can observe from Table **6**, four variables scored a coefficient of determination value greater than 0.5, and accordingly, 'mask_use', 'all_bed_capacity', 'new_cases', and 'pneumonia_st' were ranked from first to fourth, respectively and were used to build all models in boosting models. However, those feature variables with a coefficient of determination value less than 0.5 were excluded from the boosting model building.

Table **6:**

a Sensitivity analysis used to rank the inputs for boosting the model.

Features included	A longer description of feature variables	DC	Rank
mask_use	Percent of the population reporting always wearing a mask	0.867	1st
all_bed_capacity	Total number of beds that exists at the location	0.815	2nd
new_cases	Daily number of new cases	0.796	3rd
pneumonia_st	The ratio of pneumonia deaths to the average annual deaths	0.768	4th
icu_bed_capacity	Total number of ICU beds that exists at the location	0.421	5th
hosp_admission	Daily COVID-19 hospital admission	0.401	6th
daily_infection	The number of daily infections	0.253	7th

4.3. AI-driven single models

In this study, seven single AI-driven and weak-performing black-box models were developed to predict mortality due to COVID-19 in Eastern Africa for ensemble and boosting models. For ensemble modelling, four of them (ANFIS, SVM, FFNN, and MLR) were used, and for boosting model, three of them (KNN, ANN-6 and SVM) were used to predict mortality. Therefore, in this section, the result of single models for ensemble and boosting processes are presented as follows:

4.3.1. AI-driven single models for ensemble modelling

The ANFIS, SVM, FFNN, and MLR were trained and tested for each combination of input variables in the ensemble modelling process. Hence, results from each model are presented in Table **7.**

Table **7:**

Single black-box models to predict COVID-19 in ensemble modelling.

Model	Combination of Inputs	Structure	Training		Verification	
			DC	**RMSE**	**DC**	**RMSE**
FFNN	Cases, Pos_rate, vaccine, Hosp_bed	Gaussian	0.8792	0.00148	0.859	0.001412
ANFIS	Cases, Pos_rate, vaccine, Hosp_bed	4-6-1	0.9146	0.00018	0.927	0.000125
SVM	Cases, Pos_rate, vaccine, Hosp_bed	RBF	0.8650	0.00021	0.849	0.000146
MLR	Cases, Pos_rate, vaccine, Hosp_bed	4-1	0.8021	0.00012	0.796	0.000192

The FFNN was the first type of AI-driven model used in this study for ensemble modelling to predict mortality due to COVID-19. The Levenberg–Marquardt technique was used to train this model, which had four inputs and one hidden layer, to estimate COVID-19 mortality in East Africa. Identifying the optimal structure (number of hidden

neurons) of the model was a key step in obtaining the best results in the FFNN modelling. The possession of too many neurons may result in overfitting, or too few neurons may result from incorrect information.

, a trial-and-error technique was used to determine the appropriate structure of the FFNN model until the best-fit combination and performance were observed. This technique allowed us to analyze the accuracy of the numerous models trained with the variable's hidden number. As a result, the best model structure (x-y-z) with the greatest prediction outcomes was discovered to be six hidden neurons with four inputs and one hidden layer, which was noted as (4-6-1).

The second type of AI-driven model used was the ANFIS, which assumes a fuzzy notion to manage the unpredictable circumstances of complicated data of a non-linear nature. In the modelling process, Sugeno's fuzzy inference system with hybrid algorithms was used to calibrate the parameters of the membership functions (MFs). The Gaussian, triangular, and trapezoidal MFs were investigated using a trial-and-error approach to produce the best estimation result in predicting mortality due to COVID-19. As a result, the ANFIS model with "Gaussian membership functions" trained over 41 epochs offered better prediction results than the other MFs.

SVM was the third type of AI-driven model used in the modelling. The kernel of the radial basis function (RBF) was used to generate the SVM model for the combination of all the input variables. RBF was chosen because it has fewer turning parameters and performs better than sigmoidal and polynomial models (W.c. Wang et al., 2013).

Finally, the traditional MLR technique was employed to predict COVID-19 mortality and to compare the predicted result with those of the other three types of AI-driven models. The linear connection (a-b) between the one target variable and the four input variables was determined using this model and noted as (4-1).

The results of the single black-box models in Table **7** show that the ANFIS was the best-performing AI-driven model in predicting mortality due to COVID-19, with the highest DC (0.9273) and lowest RMSE (0.000125) at the verification stage. The second,

third, and fourth best models, based on their performance, were FFNN, SVM and MLR, respectively. The daily COVID-19 data are non-linear and dynamic by its nature. Hence, the non-linear AI-driven model, ANFIS, was found to be the best model for predicting the data. However, according to the calculated DCs, the MLR was the least-performing model in predicting COVID-19 mortality.

These results showed that the best models for predicting data of a non-linear and dynamic nature are the non-linear AI-driven models, such as ANFIS and FFNN, while the linear regression estimation approach was a poorly performing model in predicting the mortality due to COVID-19 in eastern Africa. According to the findings of the single black-box models, provided in Table **7,** utilizing the best-predicting model in this study (ANFIS) might improve the performance of the prediction using FFNN, SVM, and MLR by 7 percent, 8 percent, and 13 percent, respectively.

4.3.1. AI-driven single models for boosting model

In the modelling process boosting algorithms, the data were trained and tested by using three AI-driven models (KNN, SVM and ANN-6) and one boosting model (AdaBoost). Hence, the prediction performance of each model is presented in Table **8.**

Table **8:**

Single black-box models to predict COVID-19 in boosting model.

Model	Feature combinations	Model parameters	Training dataset		Test dataset	
			RMSE	**DC**	**RMSE**	**DC**
AdaBoost	**	Samme.r	1.9358	0.9449	2.0549	0.9422
KNN	**	Uniform	3.0834	0.8601	3.1858	0.8618
SVM	**	RBF	4.3482	0.7218	4.5461	0.7171
ANN-6	**	L-BFGS-B	1.9358	0.8553	3.1749	0.8629

** mask use, all_bed, number of cases, pneumonia_st

The model that we applied in this study to boost the prediction performance of COVID-19 in Ethiopia was the AdaBoost model. In this model, a variant called "AdaBoost.samme.r" was applied. This variant works with classifiers that can show output prediction probabilities. Values of DC and RMSE obtained were 0.9422 and 2.0549, respectively. This implies that the AdaBoost model was the best performer in predicting COVID-19 mortality in Ethiopia.

The second AI-driven model used to predict COVID-19 mortality was the KNN. In this model, both assumptions of weight (uniform and distance) were tried in the modelling process. However, the KNN with 'distance' weight was going to be overfitted, and the KNN with 'uniform' weight was best fitted in the prediction process. Therefore, the values of DC and RMSE for the KNN model were 0.8618 and 3.1858, respectively. Accordingly, the KNN was the third best performer model to predict COVID-19 in Ethiopia, next to the AdaBoost and the ANN-6 models.

The third AI-driven model used to predict COVID-19 was SVM. To build the SVM model using the selected dominant feature, the kernel of the radial basis function (RBF) was applied. This function was selected due to its better performance than that of the other types of functions under the SVM in predicting COVID-19 in eastern Africa (Abegaz & Etikan, 2022). As presented in Table **8**, the performance of SVM in predicting COVID-19 was reported in the form of DC and RMSE, whereby the value of the former was 0.7171, and that of the latter was 4.5461 in the test dataset. This result implies that the prediction performance of SVM was less than that of the other prediction models.

The fourth AI-driven model used was the ANN-6. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm was selected due to its proven performance even for non-smooth optimization (Mathieu et al., 2021). This implies that the ANN-6 was a good predictor for non-linear data such as the COVID-19 daily mortality. The value of DC was 0.8620, and that of the RMSE was 3.1749 in the test dataset, implying that the ANN-6 was the second-best performer algorithm to predict COVID-19 deaths in Ethiopia, next to the boosting algorithm and the first AI-driven algorithm among three single models.

4.4. The correlation analysis

The relationship between the actual and the predicted values of daily mortality due to COVID-19 using four AI-driven models (ANFIS, FFNN, SVM, and MLR) was correlated with ensemble model value in eastern Africa. In addition, three single models (KNN, ANN-6 and SVM) modelled in the boosting process were correlated with the boosting algorithm model in Ethiopia. Therefore, in this section, the result of correlation analysis for both ensemble modelling and boosting model are presented below:

4.4.1. Correlation analysis for ensemble modelling

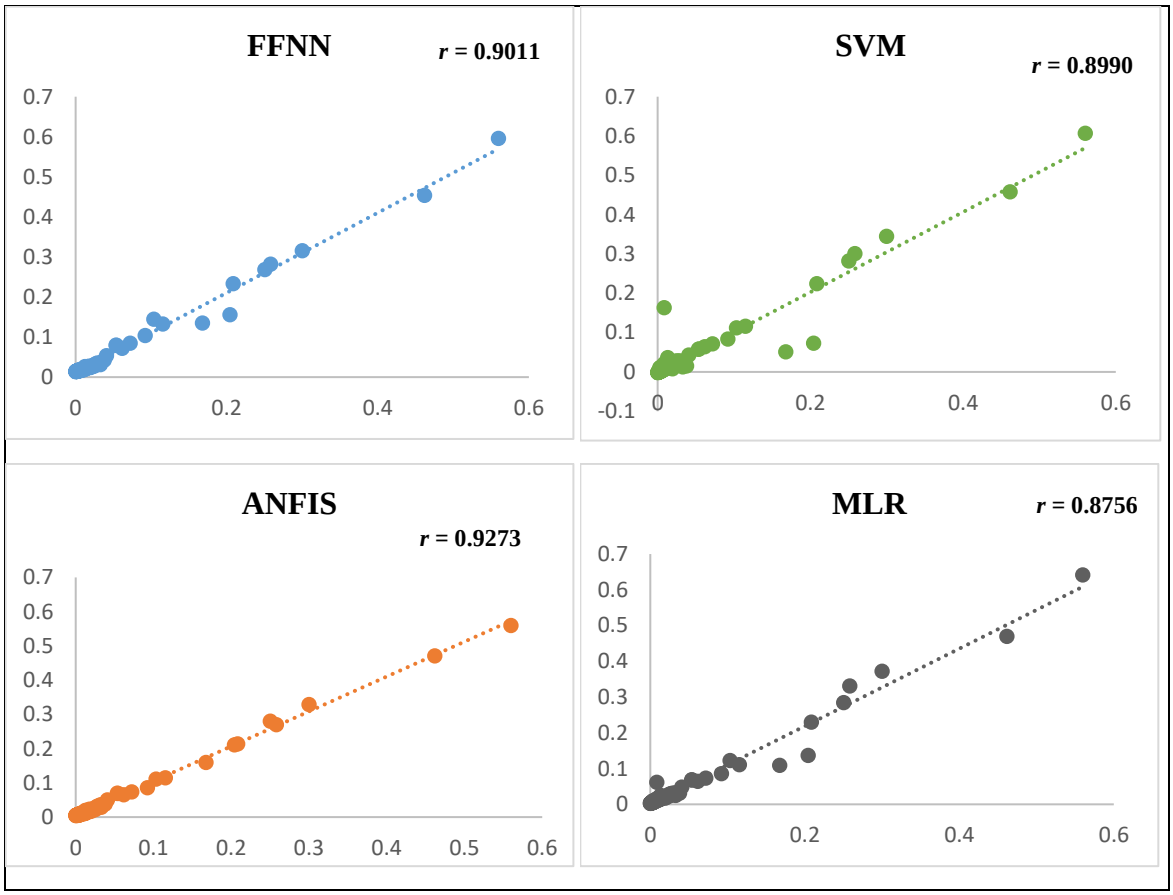
The correlation between the estimated values obtained from several AI-driven models and the observed value using a scatter plot diagram is presented in Figure **11.** In this diagram, the estimated performances of the ANFIS, FFNN, SVM, and MLR models are compared in terms of their predictions of COVID-19 in East Africa. As a result, the ANFIS model indicated fewer spread points in the linking and produced better-estimated values than the other models. This might be attributed to ANFIS's capacity to anticipate non-linear data, such as COVID-19 data, as it has a greater coefficient of determination than the other AI-driven and MLR models.

The finding from the diagram supports those of the other analyses and modellings, which showed that the non-linear predicting approaches performed better than the linear predicting approaches. Moreover, the ANFIS was the highest-performing model in predicting COVID-19 in eastern Africa.

According to a correlation analysis result, we understood that the scatter plot diagram proved that the non-linear predicting approaches performed better than the linear approaches. More specifically, among all the models, the ANFIS model was the best-performing predicting approach for the daily COVID-19 data. This finding is similar to the finding of a study conducted on daily suspended sediment load data using the AI-driven ensemble model (Nourani, Gokcekus, & Gelete, 2021).

Figure **11**:

Correlation between actual and predicted COVID-19 mortality during ensemble model



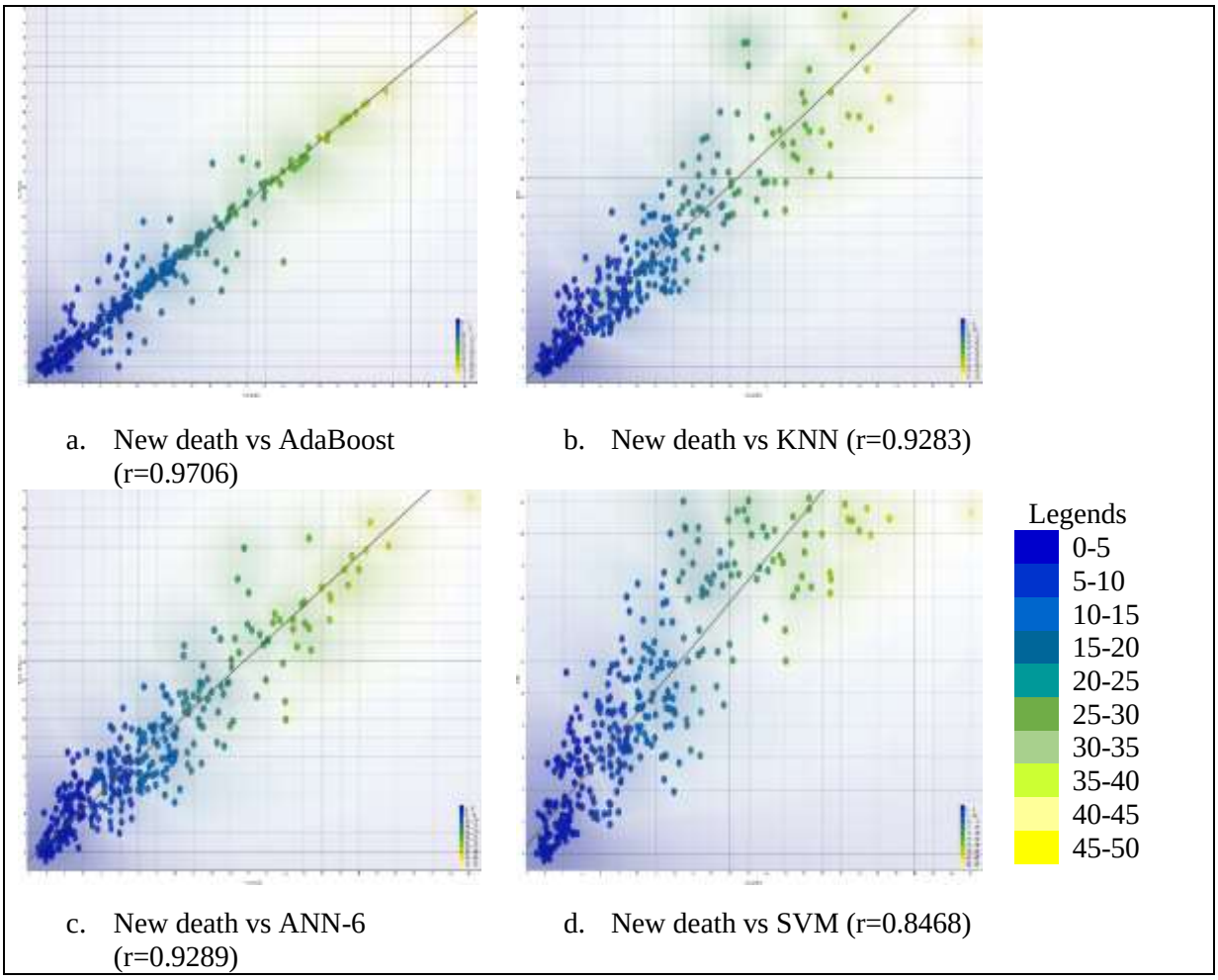
4.4.2. Correlation analysis for boosting model

The relationship between the actual and the predicted values of daily mortality due to COVID-19 using four AI-driven models (AdaBoost, KNN, SVM and ANN-6) was calculated and presented in Figure **12**. In this visual presentation, the rank of AI-driven models in predicting COVID-19 mortality was presented in bivariate correlation values. Hence, the correlation values were 0.9706, 0.9289, 0.9283, and 0.8468 for AdaBoost, ANN-6, KNN, and SVM, respectively. This implies that AdaBoost, ANN-6, KNN, and SVM were the first, second, third and fourth models, respectively, to indicate fewer spread

points in the correlation with mortality due to COVID-19 in Ethiopia and thereby produce a better-estimated value of the mortality.

Figure **12:**

Correlation between actual and predicted COVID-19 mortality during boosting model



a. New death vs AdaBoost (r=0.9706)

b. New death vs KNN (r=0.9283)

c. New death vs ANN-6 (r=0.9289)

d. New death vs SVM (r=0.8468)

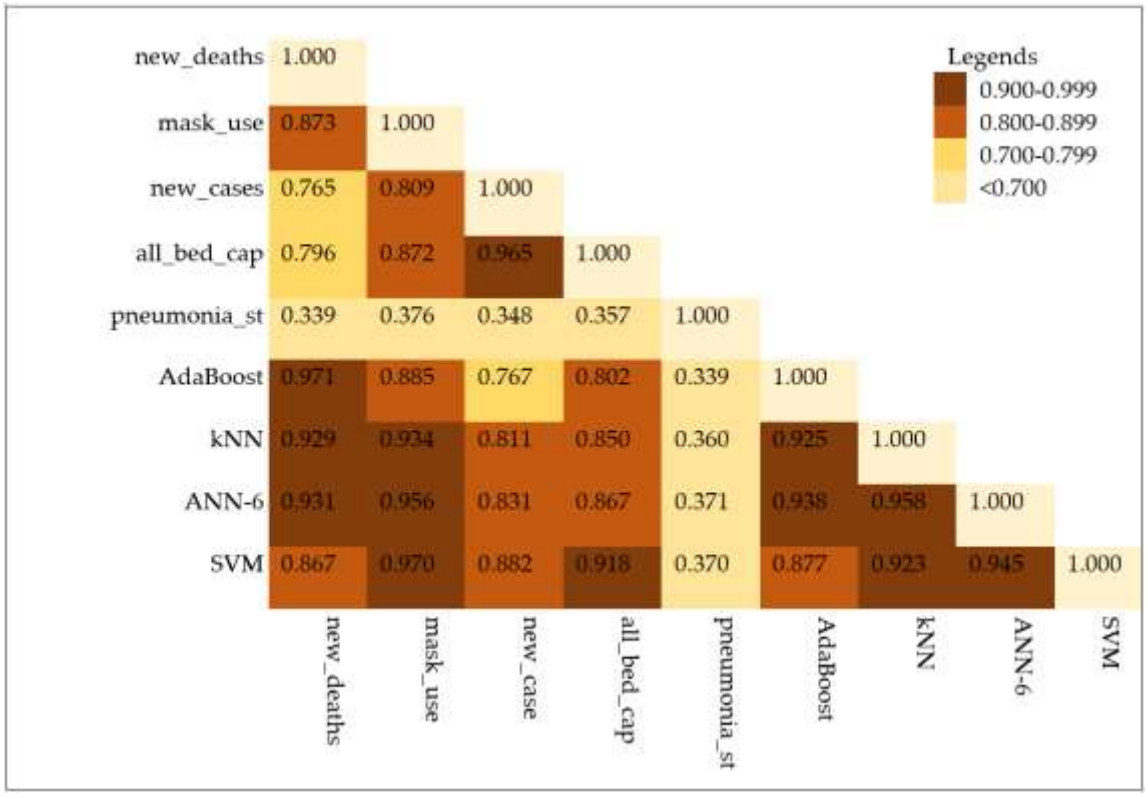
The scatter plot in Figure **11** and the results of the AI-driven models' analysis in Table **7** helped us understand that the boosting algorithm outperformed others in terms of predicting COVID-19 mortality in Ethiopia. According to this finding, the AdaBoost algorithm was the most effective AI-driven model for predicting COVID-19 data that was gathered daily.

A bivariate correlation analysis using the Spearman correlation coefficient was conducted, and the result is presented in Figure **13**. In this analysis, the observed value of daily mortality was correlated with each observed feature variable, and each predicted value from the AI-driven models (AdaBoost, KNN, ANN-6 and SVM).

The predicted values with AdaBoost, ANN-6, KNN and SVM algorithms were the first, second, third and fourth highly correlated ones with values of 0.971, 0.931, 0.929, and 0.867, respectively. In addition to this, mask use, all bed capacity, and daily new cases were the first three highly correlated feature variables with values of 0.873, 0.796, and 0.765, respectively. However, the pneumonia case was the lowest correlated feature variable. Hence, we understood from this result that the Spearman correlation value was improved among AI-driven models.

Figure **13**:

Correlation statistics among the input variables and the predicted mortality



4.6. Taylor's diagram for model comparison

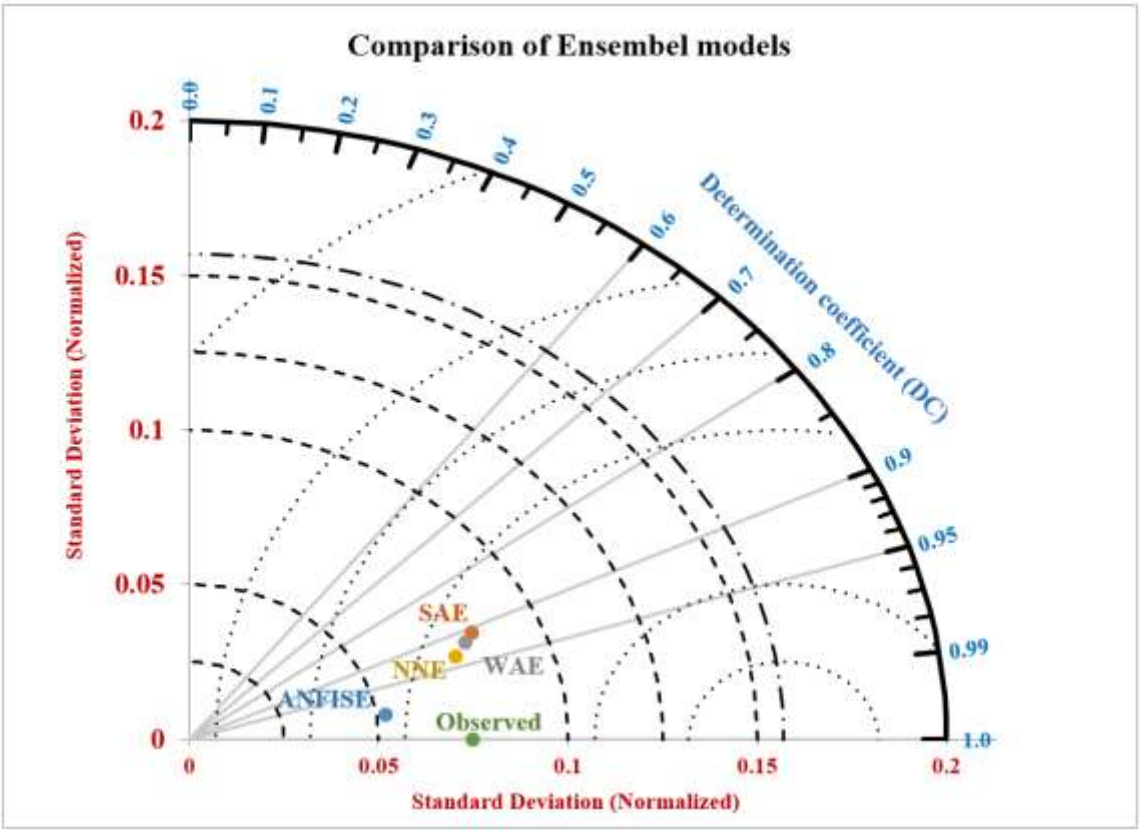
To summarize the performance of multiple models in a single diagram, Taylor's diagram was constructed for both ensemble and boosting models separately.

4.6.1. Taylor's Diagram for ensemble modelling

According to Taylor K.E., we can summarize multiple models' performances in a single diagram that can help us to easily visualize and understand which model performs better (Taylor, 2001). Therefore, four ensemble approaches were assessed using a two-dimensional Taylor diagram, as presented in Figure **14**, which coordinates the correlation coefficients (r) and the standard deviations (SD) for both the observed and predicted values of the ensemble models (ANFISE, NNE, WAE, and SAE).

Figure **14**:

Performance of ensemble models using a normalized Taylor diagram.



The advantage of using the Taylor diagram is that it combines the predicted performances of different models in a single visual display that quantifies the level of resemblance between the observed and the predicted values. It is observed from Figure **14** that the ANFISE was the best approach in predicting COVID-19 in the eastern Africa region, with (r = 0.9852 and SD = 0.0523), and the SAE was the poorest-performing ensemble approach, with (r = 0.9073 and SD = 0.0821).

In addition to the statistical evidence, the Taylor diagram showed that the 'r' Vs 'SD' coordinate of the ANFIS ensemble approach was closer to the observed value than the rest of the ensemble approaches, and we can see that the coordinate for the SAE was far from the observed value compared to the other ensemble approaches. This closeness showed that the predicted values obtained from the ANFISE were more closely related to the observed value. Hence, this proves that this ensemble model has the best prediction capability among the other models.

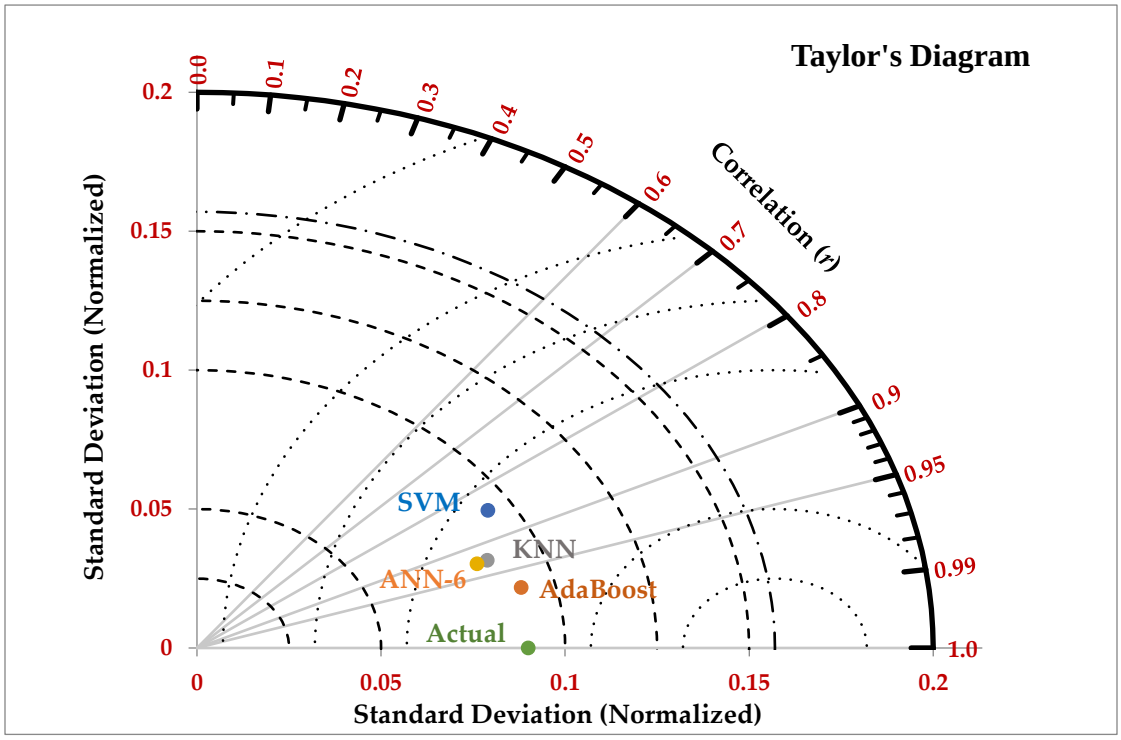
4.6.2. Taylor's diagram for boosting model

We visualized the performance of various AI-driven models in a single diagram, as shown in Figure **15**, which is a two-dimensional diagram that coordinates the standard deviation (SD) and correlation coefficient (r) of each AI-driven model's predicted value (using AdaBoost, KNN, ANN-6, and SVM) and the observed values of COVID-19 mortality in Ethiopia. The significance of using this diagram is that it quantifies the degree of similarity between the predicted values and the observed values of mortality while simultaneously displaying the predicted performance of various models in a single visual display.

Figure **15** makes it clear that "AdaBoost" was the AI-driven model that performed the best in predicting COVID-19 mortality in Ethiopia, with (r=0.9706 and SD of 0.0907), and that the SVM was the model that performed the worst, with (r=0.8468 and SD=0.0934).

Figure **15**:

Taylor diagram showing the prediction performance of models



CHAPTER V

DISCUSSION

In this chapter, the research paper tried to briefly discuss and present the summary findings, implications of the study, limitations and future directions of the study, and contribution of the study to the field.

5.1. Summary of Findings

The purpose of this work was to construct and assess ensemble and boosting models powered by artificial intelligence (AI), such as AdaBoost, KNN, ANFIS, SVM, FFNN, ANN-6, and MLR, for predicting COVID-19 mortality in Eastern Africa. The results shed important light on how AI approaches may improve the precision of death predictions and guide regional public health policies. These models successfully illustrated the importance of multiple variables, including Bed capacity, mask use, number of cases, pneumonia status, positive rate, and vaccination status, in predicting COVID-19 mortality by capturing the intricate interactions between them.

The ensemble learning approaches showed improved performance, leading to increased accuracy and resilience by pooling the predictions from many models. Overall, by showcasing the potential of AI approaches to deliver precise and trustworthy forecasts for COVID-19 mortality, this research advances the fields of public health and data mining and knowledge discovery.

5.2. Interpretation of results

Several important determinants of COVID-19 mortality in Eastern Africa were found after data analysis. The key drivers of death rates were found to be positive rates, hospital beds/1000, new cases and the number of vaccinated individuals, mask use and pneumonia status of patients. The AI-driven models accurately reflected these correlations and offered information on the relative weights of various predictors. The impact of the pandemic on the region's most vulnerable groups can be lessened with the use of targeted actions and resource allocation plans informed by this knowledge. The models also

demonstrated the promise of AI methods for dealing with dynamic and complicated information.

5.3. Ensemble Models

The AI-driven ensemble model was developed using the estimated outputs from three single AI-driven models (ANFIS, FFNN, and SVM) and one classical regression model (MLR) as the input variables for the ensemble modelling. This model was developed to boost the efficiency of the single models in terms of prediction capability. Four ensemble approaches (SAE, WAE, ANFISE, and NNE), as novel ensemble processes, were developed to predict COVID-19 in East Africa, and the results are presented in Table **9**. Accordingly, the (a-b) structure for the SAE was applied to display the numbers of outputs and inputs used for the prediction. The structure (a, b, c, d) was the structure for the WAE that denoted the weights of the FFNN, ANFIS, SVM, and MLR single models, respectively.

Table **9**:

An ensemble approach used to model COVID-19 mortality in eastern Africa.

Ensemble Method	Selected Structure	Calibration		Verification	
		DC	**RMSE**	**DC**	**RMSE**
SAE	3-1	0.9446	0.000821	0.9073	0.000245
WAE	0.243, 0.269, 0.249, 0.22	0.9250	0.000123	0.9190	0.000156
ANFIS_E	Gaussian 3	0.9292	0.001658	0.9886	0.000012
NNE	3-6-2	0.9286	0.000120	0.9356	0.000132

The ANFISE was best-performing among all the ensemble model development combinations. This is due to its resilience in integrating both the fuzzy concept and the artificial neural network capability, which provided the present ANFIS framework. The NNE, WAE, and SAE were the second, third, and fourth-best predictors of COVID-19. The weighted ensemble technique outperformed the simple average ensemble approach. This is because the WAE assigns weights to parameters depending on their relevance.

The Levenberg–Marquardt method was used to train the NNE model, as it applied to the FFNN, and the tangent sigmoid activation function was utilized for the hidden and output layers. The study conducted by Sahoo et al. (Sahoo, Ray, & Wade, 2005) indicated that the FFNN approach has the fastest convergence ability; hence, it was used more often in this study than the other ANN training techniques. We used a trial-and-error method to determine the correct number of hidden layers and the optimal epoch number. In ANFISE, Sugeno's fuzzy inference system, using a hybrid training approach, was used to calibrate the membership function parameters comparable to those of the ANFIS single black-box model. As a result, the ANFISE greatly improved the accuracy of the single models.

5.4. Comparison of models

5.4.1. Comparison of ensemble modelling with a single AI-driven model

The comparison of the prediction performances of the ensemble models and single AI-driven models at the verification and training phases is presented in Table **10**. In this table, the NNE boosted the predicting performance of the single models FFNN, ANFIS, SVM, and MLR by 5.6, 2.1, 7.1, and 13.4 percent, respectively. In addition to this, the ANFISE boosted the performance of FFNN, ANFIS, SVM, and MLR by 13, 6.1, 13.9, and 19.3 percent, respectively. These numbers show that the capacity for the prediction of COVID-19 was increased in the case of the ensemble models rather than the single models, and these findings were compared to the findings of studies conducted in different fields using AI ensemble models (Kazienko et al., 2013; Y. Wang et al., 2021; Yaşar et al., 2021).

Hence, these findings showed that ensemble models could be applied to the prediction of COVID-19 in the eastern Africa region more effectively than single AI-driven models. In addition to this, the findings prove that the non-linear ensemble models are more capable than the linear ensemble models. This might be due to the incapability of linear ensemble approaches to undergo another black-box learning process, unlike the non-linear approaches.

Table **10:**

Comparison of prediction levels of single AI models and ensemble models.

Ensemble Models	**Single Models**	**Ensemble vs Single Models**	The Difference in Percent (%)	
			Verification	**Training**
NNE	FFNN	NNE vs FFNN	5.6%	4.9%
	ANFIS	NNE vs. ANFIS	2.1%	1.4%
	SVM	NNE vs. SVM	7.1%	6.4%
	MLR	NNE vs. MLR	13.4%	12.7%
ANFIS_E	FFNN	ANFIS_E vs. FFNN	13%	5%
	ANFIS	ANFIS_E vs. ANFIS	6.1%	1.4%
	SVM	ANFIS_E vs. SVM	13.9%	6.4%
	MLR	ANFIS_E vs MLR	19.3%	12.7%

5.4.2. Comparison of AdaBoost with a single AI-driven model

Table **11** compares the boosting model with AI-driven models in terms of prediction performance across training and test datasets. In a training dataset, the AdaBoost model improved the prediction accuracy of KNN, SVM, and ANN-6 models by 8.48 percent, 22.31 percent, and 8.96 percent, respectively.

Additionally, it improved the accuracy of prediction for KNN, SVM, and ANN-6 models in a test dataset by 7.94 percent, 22.51 percent, and 8.02 percent, respectively. The results indicated that ensemble boosting models could be used to predict COVID-19 mortality in Ethiopia more effectively than the tested individual AI-driven weak-performing models.

Table **11**:

Comparison of boosting model with weak AI-driven models

Boosted Model vs Single Model	**Difference in percentage**	
	Training dataset	**Test dataset**
AdaBoost vs KNN	8.48%	7.94%
AdaBoost vs SVM	22.31%	22.51%
AdaBoost vs ANN-6	8.96%	8.02%
kNN vs. SVM	13.83%	14.57%
kNN vs. ANN-6	0.48%	0.08%
ANN-6 vs SVM	13.35%	14.49%

5.5. Implication of the study

The results of this study have important ramifications for Eastern African public health practice and policy. The AI-driven models created in this study can help with resource allocation and planning by identifying those who are more likely to die from COVID-19. Because of this, healthcare professionals may prioritise patient care and spend resources appropriately. In order to lessen the impact of the disease, policymakers can use these models to guide targeted initiatives like vaccination drives and preventative measures. However, it is critical to recognize the study's constraints, notably its dependence on reliable and accessible data.

In order to increase the usefulness and generalizability of these models, future research should concentrate on verifying and improving them using bigger and more varied datasets. Overall, the identified predictors of death, such as age, comorbidities, and the state of the healthcare system, highlight the need for specialized assistance, improved healthcare systems, and sufficient funding to handle the COVID-19 pandemic in Eastern Africa.

5.6. Future directions of the study

Despite the encouraging findings, there are several limitations to this study that need to be noted. First, the quantity and calibre of data were crucial to the precision of our prediction models. The effectiveness of the models may have been impacted by issues including constrained testing capacity, differences in reporting standards, and data gaps. Therefore, to increase the precision and dependability of the predictions, future research should concentrate on developing data-gathering methods and tackling data quality challenges.

Furthermore, because this study's models were created specifically for Eastern Africa, they might not be readily transferable to other locations with varied demographic, epidemiological, and healthcare features. In order to increase the generalizability of these models, more research is required to test and improve them using a variety of datasets from various geographical regions.

5.7. Contribution to the filed

This work contributes to the field of COVID-19-specific AI-driven death prediction in Eastern Africa. This study advances knowledge of AI applications in public health and epidemiology by proving the efficacy of ensemble and boosting models. The findings give an understanding of the variables affecting COVID-19 mortality and a foundation for creating forecasting models that might help with decision-making in environments with constrained resources.

The use of AI methods in mortality prediction has the potential to improve public health initiatives, resource allocation, and epidemic response planning. Additionally, by offering light on the particular difficulties and factors in the area, this study adds to the body of knowledge on COVID-19 in Eastern Africa.

In summary, the study's promise for forecasting COVID-19 mortality in Eastern Africa is demonstrated by the ensemble and boosting models that are AI-driven. The findings highlight the significance of age, comorbidities, and healthcare systems as major factors in death rates. However, it is imperative to address the research's shortcomings,

such as the requirement for validation in a variety of demographics and data quality difficulties. Overall, applying AI approaches to mortality prediction can significantly improve public health responses, resource allocation, and decision-making processes to successfully address the COVID-19 epidemic in Eastern Africa and beyond.

CHAPTER VI

CONCLUSION AND RECOMMENDATIONS

In this work, the ability of AI-driven ensemble and boosting models to predict mortality due to COVID-19 in East Africa was investigated. Before predicting COVID-19 mortality using models, the data were normalized, and a sensitivity analysis was performed to identify the best dominant input variables.

Among four single AI-driven models used in the ensemble, ANFIS outperformed the other models due to its ability to analyze non-linear, dynamic, and complicated processes using the fuzzy concept and neural network idea. Four ensemble techniques were modelled to improve the performance of the single AI-driven models by aggregating the results from each AI-driven model and using the aggregated result as an input for the ensemble modelling.

Because of their capacity to handle unpredictable, non-stationarity, and complicated data, the non-linear ensemble techniques (ANFISE and NNE) outperformed the linear ensemble approaches (SAE and WAE). ANFISE was the best-performing ensemble technique, improving the prediction performance of the single AI-driven models by 13, 6.1, 13.9, and 19.3 percent, respectively.

Finally, three single AI-driven models and one boosting model were developed to predict mortality, and the prediction performance of these three models was compared with that of the boosting model. At the verification stage using the testing dataset, AdaBoost boosted the prediction of the performance of three models, KNN, SVM, and ANN-6 models, in a testing dataset by 7.94, 22.51, and 8.02 percent, respectively.

Overall, the outcome of this study demonstrated the potential capacity of ensemble and boosting models to predict mortality due to COVID-19. The result obtained from the ANFIS ensemble model and the Boosting algorithms demonstrated that aggregating the outputs of separate AI-driven models leads to a better prediction than employing them individually.

The study's weakness was that it only used black-box models to calculate COVID-19 mortality. As a result, the use of physically based models in the assembly process should be investigated in future research. Furthermore, this study used only two years of daily recorded COVID-19 mortality and other feature variables to develop the single models and the boosting model. Therefore, it is important to test these AI-driven boosting models for further data with a large number of observations in future studies.

REFERENCES

Abba, S. I., Linh, N. T. T., Abdullahi, J., Ali, S. I. A., Pham, Q. B., Abdulkadir, R. A., . . . Anh, D. T. (2020). Hybrid machine learning ensemble techniques for modeling dissolved oxygen concentration. *IEEE Access, 8*, 157218-157237.

Abdunabi, T. A. (2016). A framework for ensemble predictive modeling.

Abegaz, K. H., & Atomssa, E. M. (2017). Data mining of access to tetanus toxoid immunization among women of childbearing age in Ethiopia. *Machine Learning Research, 2*(2), 54-60.

Abegaz, K. H., & Etikan, İ. (2022). Artificial Intelligence-Driven Ensemble Model for Predicting Mortality Due to COVID-19 in East Africa. *Diagnostics, 12*(11), 2861.

Abegaz, K. H., & Habtewold, E. M. (2019). Trend and barriers of antenatal care utilization from 2000 to 2016 Ethiopian DHS: a data mining approach. *Scientific African, 3*, e00063.

Ajami, N. K., Duan, Q., Gao, X., & Sorooshian, S. (2006). Multimodel combination techniques for analysis of hydrological simulations: Application to distributed model intercomparison project results. *Journal of Hydrometeorology, 7*(4), 755-768.

Arora, N., Banerjee, A. K., & Narasu, M. L. (2020). The Role of artificial intelligence in Tackling COVID-19. In (Vol. 15, pp. 717-724): Future Medicine.

Ayalew, A. M., Salau, A. O., Abeje, B. T., & Enyew, B. (2022). Detection and classification of COVID-19 disease from X-ray images using convolutional neural networks and histogram of oriented gradients. *Biomedical Signal Processing and Control, 74*, 103530.

Baik, S.-M., Lee, M., Hong, K.-S., & Park, D.-J. (2022). Development of Machine-Learning Model to Predict COVID-19 Mortality: Application of Ensemble Model and Regarding Feature Impacts. *Diagnostics, 12*(6), 1464.

Bitew, F. H., Nyarko, S. H., Potter, L., & Sparks, C. S. (2020). Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey. *Genus, 76*, 1-16.

BogoToBogo. (2023). Artificial Neural Network (ANN) 6 Training via BFGS. Retrieved January 2023

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123-140.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32.

Brilliant.org. (2022). Feedforward Nural Network. Retrieved August 2021, from https://brilliant.org/wiki/feedforward-neural-networks

Chowdhury, D., Banerjee, S., Sannigrahi, M., Chakraborty, A., Das, A., Dey, A., & Dwivedi, A. D. (2022). Federated learning-based Covid-19 detection. *Expert Systems*, e13173.

Cui, S., Wang, Y., Wang, D., Sai, Q., Huang, Z., & Cheng, T. (2021). A two-layer nested, heterogeneous ensemble learning predictive method for COVID-19 mortality. *Applied Soft Computing, 113*, 107946.

Dawson, C. W., Abrahart, R. J., & See, L. M. (2007). HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling & Software, 22*(7), 1034-1052.

Dejene, B. E., Abuhay, T. M., & Bogale, D. S. (2022). Predicting the level of anaemia among Ethiopian pregnant women using homogeneous ensemble machine learning algorithm. *BMC Medical Informatics and Decision Making, 22*(1), 1-11.

Dietterich, T. G. (2000). *Ensemble methods in machine learning.* Paper presented at the Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1.

Dong, J., Wu, H., Zhou, D., Li, K., Zhang, Y., Ji, H., . . . Liu, Z. (2021). Application of big data and artificial intelligence in COVID-19 prevention, diagnosis, treatment and management decisions in China. *Journal of Medical Systems, 45*(9), 84.

Edeh, M. O., Dalal, S., Dhaou, I. B., Agubosim, C. C., Umoke, C. C., Richard-Nnabu, N. E., & Dahiya, N. (2022). Artificial Intelligence-based ensemble learning model for prediction of hepatitis C disease. *Frontiers in Public Health*, 847.

Erdaw, Y., & Tachbele, E. (2021). Machine learning model applied on chest X-ray images enables automatic detection of COVID-19 cases with high accuracy. *International Journal of General Medicine*, 4923-4931.

Fernandes, A. M., Oliveira, P., Moura, J. P., Oliveira, A. A., Falco, V., Correia, M. J., & Melo-Pinto, P. (2011). Determination of anthocyanin concentration in whole grape skins using hyperspectral imaging and adaptive boosting neural networks. *Journal of Food Engineering, 105*(2), 216-226.

Freund, Y., & Schapire, R. E. (1996). *Experiments with a new boosting algorithm.* Paper presented at the icml.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and system sciences, 55*(1), 119-139.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence, 14*(771-780), 1612.

Gao, Y., Cai, G.-Y., Fang, W., Li, H.-Y., Wang, S.-Y., Chen, L., . . . Cui, P.-F. (2020). Machine learning-based early warning system enables accurate mortality risk prediction for COVID-19. *Nature communications, 11*(1), 5033.

Gao, Y., Chen, L., Chi, J., Zeng, S., Feng, X., Li, H., . . . Wang, Y. (2021). Development and validation of an online model to predict critical COVID-19 with immune-inflammatory parameters. *Journal of Intensive Care, 9*(1), 1-12.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R*: Springer.

GAVI. (2022). the vaccine work, does COVID-19 compare to past pandemics

Guo, Q., & He, Z. (2021). Prediction of the confirmed cases and deaths of global COVID-19 using artificial intelligence. *Environmental Science and Pollution Research, 28*, 11672-11682.

Hasell, J., Mathieu, E., Beltekian, D., Macdonald, B., Giattino, C., Ortiz-Ospina, E., . . . Ritchie, H. (2020). A cross-country database of COVID-19 testing. *Scientific data, 7*(1), 345.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on pattern analysis and machine intelligence, 20*(8), 832-844.

Hu, Z., Ge, Q., Li, S., Jin, L., & Xiong, M. (2020). Artificial intelligence forecasting of covid-19 in China. *arXiv preprint arXiv:2002.07112*.

ITA. (2023). Ethiopia-Country commercial guide. Retrieved December 2022

Jang, J.-S. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics, 23*(3), 665-685.

Jang, J.-S. R., Sun, C.-T., & Mizutani, E. (1997). Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book Review]. *IEEE Transactions on automatic control, 42*(10), 1482-1484.

Kalteh, A. M. (2013). Monthly river flow forecasting using artificial neural network and support vector regression models coupled with wavelet transform. *Computers & Geosciences, 54*, 1-8.

Karaarslan, E., & Aydın, D. (2021). An artificial intelligence–based decision support and resource management system for the COVID-19 pandemic. In *Data Science for COVID-19* (pp. 25-49): Elsevier.

Kazienko, P., Lughofer, E., & Trawiński, B. (2013). Hybrid and ensemble methods in machine learning J. UCS special issue. *J Univers Comput Sci, 19*(4), 457-461.

Ko, H., Chung, H., Kang, W. S., Park, C., Kim, D. W., Kim, S. E., . . . Seo, J. H. (2020). An artificial intelligence model to predict the mortality of COVID-19 patients at hospital admission time using routine blood samples: development and validation of an ensemble model. *Journal of medical Internet research, 22*(12), e25442.

Kolozsvári, L. R., Bérczes, T., Hajdu, A., Gesztelyi, R., Tiba, A., Varga, I., . . . Garbóczy, S. (2021). Predicting the epidemic curve of the coronavirus (SARS-CoV-2) disease (COVID-19) using artificial intelligence: An application on the first and second waves. *Informatics in Medicine Unlocked, 25*(100691).

Kumar, K. (2021). Machine Learning-Based Ensemble Approach for Predicting the Mortality Risk of COVID-19 Patients: A Case Study. In *Intelligent Data Analysis for COVID-19 Pandemic* (pp. 1-25): Springer.

Kuncheva, L. I. (2014). *Combining pattern classifiers: methods and algorithms*: John Wiley & Sons.

Liu, H., Tian, H.-q., Li, Y.-f., & Zhang, L. (2015). Comparison of four Adaboost algorithm-based artificial neural networks in wind speed predictions. *Energy Conversion and Management, 92*, 67-81.

Lou, L., Xia, W., Sun, Z., Quan, S., Yin, S., Gao, Z., & Lin, C. (2023). COVID-19 mortality prediction using ensemble learning and grey wolf optimization. *PeerJ Computer Science, 9*, e1209.

Mamdani, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies, 7*(1), 1-13.

Mariam, B. G., & Mariam, T. H. (2015). Application of data mining techniques for predicting CD4 status of patients on ART in Jimma and Bonga Hospitals, Ethiopia. *Journal of Health & Medical Informatics, 6*(6), 1-9.

Markos, Z., Doyore, F., Yifiru, M., & Haidar, J. (2014). Predicting Under nutrition status of under-five children using data mining techniques: The Case of 2011 Ethiopian Demographic and Health Survey. *J Health Med Inform*, *5*(2).

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., . . . Rodés-Guirao, L. (2021). A global database of COVID-19 vaccinations. *Nature human behaviour, 5*(7), 947-953.

Min, H., & Luo, X. (2016). Calibration of soft sensor by using Just-in-time modelling and AdaBoost learning method. *Chinese Journal of chemical engineering, 24*(8), 1038-1046.

Nourani, V., Elkiran, G., & Abba, S. (2018). Wastewater treatment plant performance analysis using artificial intelligence–an ensemble approach. *Water Science and Technology, 78*(10), 2064-2076.

Nourani, V., Gokcekus, H., & Gelete, G. (2021). Estimation of suspended sediment load using artificial intelligence-based ensemble model. *Complexity, 2021*, 1-19.

Nourani, V., Gökçekuş, H., & Umar, I. K. (2020). Artificial intelligence-based ensemble model for prediction of vehicular traffic noise. *Environmental Research, 180*, 108852.

Rojas, R. (2009). AdaBoost and the super bowl of Classifiers a tutorial introduction to adaptive boosting. *Freie University, Berlin, Tech. Rep*.

Sahle, G. (2016). Ethiopic maternal care data mining: discovering the factors that affect postnatal care visit in Ethiopia. *Health information science and systems, 4*, 1-8.

Sahoo, G., Ray, C., & Wade, H. (2005). Pesticide prediction in groundwater in North Carolina domestic wells using artificial neural networks. *Ecological Modelling, 183*(1), 29-46.

Schapire, R. E. (2013). Explaining AdaBoost. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, 37-52.

Schiaffino, S., Codari, M., Cozzi, A., Albano, D., Alì, M., Arioli, R., . . . Carriero, S. (2021). Machine learning to predict in-hospital mortality in covid-19 patients using computed tomography-derived pulmonary and vascular features. *Journal of Personalized Medicine, 11*(6), 501.

Sharghi, E., Nourani, V., & Behfar, N. (2018). Earthfill dam seepage analysis using ensemble artificial intelligence-based modelling. *Journal of Hydroinformatics, 20*(5), 1071-1084.

Statista. (2022). African Country with the Largest Population as of 2020.

Sun, J., Fujita, H., Chen, P., & Li, H. (2017). Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble. *Knowledge-Based Systems, 120*, 4-14.

Sun, W., & Gao, Q. (2019). Exploration of energy saving potential in China power industry based on Adaboost back propagation neural network. *Journal of Cleaner Production, 217*, 257-266.

Taherkhani, A., Cosma, G., & McGinnity, T. M. (2020). AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing, 404*, 351-366.

Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modelling and control. *IEEE Transactions on Systems, Man, and Cybernetics* (1), 116-132.

Tanty, R., & Desmukh, T. S. (2015). Application of artificial neural network in hydrology—A review. *Int. J. Eng. Technol. Res, 4*, 184-188.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of geophysical research: atmospheres, 106*(D7), 7183-7192.

Tsukamoto, Y. (1979). An approach to fuzzy reasoning method. *Advances in fuzzy set theory and applications*.

Ullah, F., Moon, J., Naeem, H., & Jabbar, S. (2022). Explainable artificial intelligence approach in combating real-time surveillance of the COVID-19 pandemic from CT scan and X-ray images using ensemble model. *The Journal of Supercomputing, 78*(17), 19246-19271.

UN. ECA. (2022). The Economic and Social Impacts of the COVID-19 Crisis on Eastern Africa: Strategies for Building Back Better. Retrieved November 2020

USAID. (2022). East Africa Regional. Global Health.

Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14*(4), 337-339.

Wang, W.-c., Xu, D.-m., Chau, K.-w., & Chen, S. (2013). Improved annual rainfall-runoff forecasting using PSO–SVM model based on EEMD. *Journal of Hydroinformatics, 15*(4), 1377-1390.

Wang, Y., Xu, C., Yao, S., Wang, L., Zhao, Y., Ren, J., & Li, Y. (2021). Estimating the COVID-19 prevalence and mortality using a novel data-driven hybrid model based on ensemble empirical mode decomposition. *Scientific Reports, 11*(1), 21413.

Web, R. (2022). East & Horn of Africa COVID-19 Situation Report—#44, 10 March 2021 Update. 2021. . Retrieved October 2022

WHO. (2020). Coronavirus (COVID-19) Dashboard; World Health Organization: Geneva, Switzerland. Retrieved January 2023

World Atlas. (2022). East African Countries. May 2021.

Wu, C., Hwang, M., Huang, T.-H., Chen, Y.-M. J., Chang, Y.-J., Ho, T.-H., . . . Ho, W.-H. (2021). Application of artificial intelligence ensemble learning model in early prediction of atrial fibrillation. *BMC Bioinformatics, 22*, 1-12.

Yaşar, Ş., Çolak, C., & Yoloğlu, S. (2021). Artificial intelligence-based prediction of Covid-19 severity on the results of protein profiling. *Computer Methods and Programs in Biomedicine, 202*, 105996.

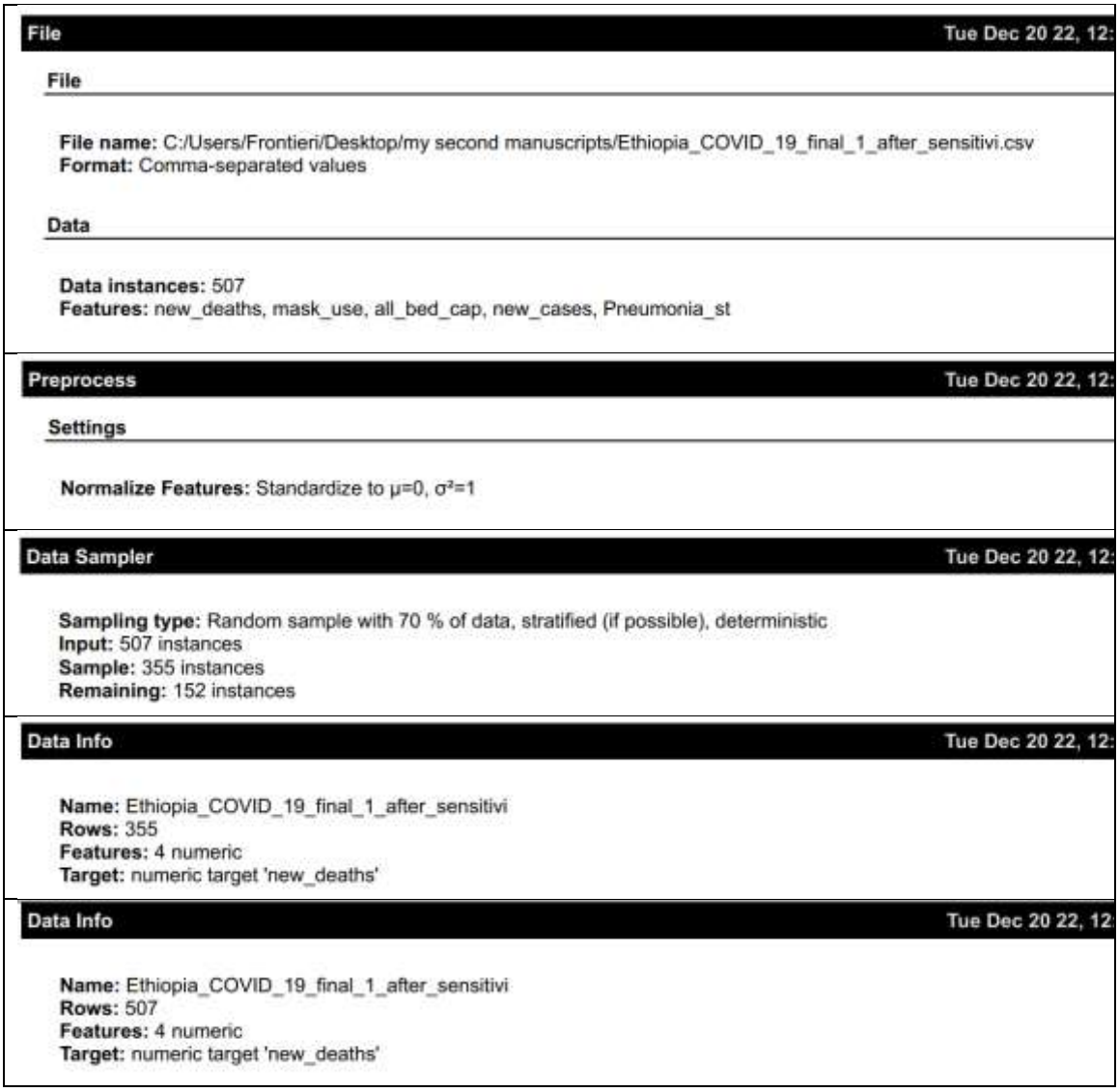
Zhao, Y., Gong, L., Zhou, B., Huang, Y., & Liu, C. (2016). Detecting tomatoes in greenhouse scenes by combining AdaBoost classifier and colour analysis. *Biosystems Engineering, 148*, 127-137.

Zhou, H., Li, Y., Zhang, Q., Xu, H., & Su, Y. (2022). Soft-sensing of effluent total phosphorus using adaptive recurrent fuzzy neural network with Gustafson-Kessel clustering. *Expert Systems with Applications, 203*, 117589.

Zhou, H., Zhang, Y., Duan, W., & Zhao, H. (2020). Nonlinear systems modelling based on self-organizing fuzzy neural network with hierarchical pruning scheme. *Applied Soft Computing, 95,* 106516.

APPENDICES

Appendix A: The Orange data mining output



AdaBoost Tue Dec 20 22, 12:

Name: AdaBoost (the Ensemble method)

Model parameters

Base estimator: tree
Number of estimators: 4
Algorithm (classification): Samme.r
Loss (regression): Square

Data

Data instances: 507
Features: mask_use, all_bed_cap, new_cases, Pneumonia_st
Target: new_deaths

kNN Tue Dec 20 22, 12:

Name: kNN

Model parameters

Number of neighbours: 2
Metric: Manhattan
Weight: Uniform

Data

Data instances: 507
Features: mask_use, all_bed_cap, new_cases, Pneumonia_st
Target: new_deaths

SVM Tue Dec 20 22, 12:

Name: SVM

Model parameters

SVM type: SVM, C=1.0, ε=0.10000000000000003
Kernel: RBF, exp(-auto|x-y|²)
Numerical tolerance: 0.001
Iteration limt: 300

Data

Data instances: 507
Features: mask_use, all_bed_cap, new_cases, Pneumonia_st
Target: new_deaths

Neural Network — Tue Dec 20 22, 12:

Name: Neural Network

Model parameters

Hidden layers: 200
Activation: tanh
Solver: L-BFGS-B
Alpha: 1
Max iterations: 500
Replicable training: True

Data

Data instances: 507
Features: mask_use, all_bed_cap, new_cases, Pneumonia_st
Target: new_deaths

Correlations — Tue Dec 20 22, 12

Spearman correlation

1	+0.885	mask_use	new_deaths
2	+0.824	all_bed_cap	new_deaths
3	+0.799	new_cases	new_deaths
4	-0.710	Pneumonia_st	new_deaths

Test and Score — Tue Dec 20 22, 12

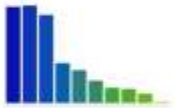
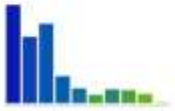
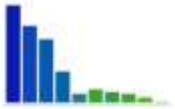
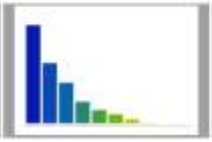
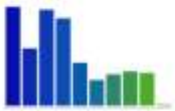
Settings

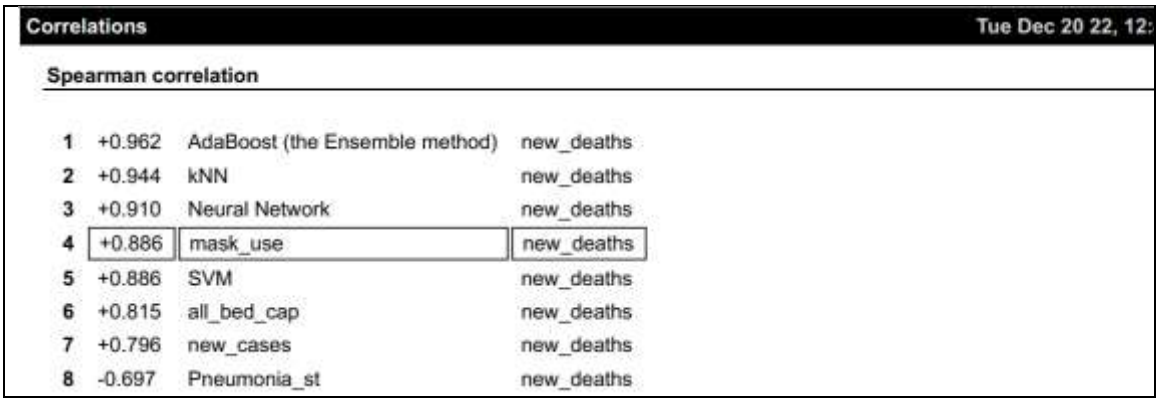
Sampling type: No sampling, test on testing data

Scores

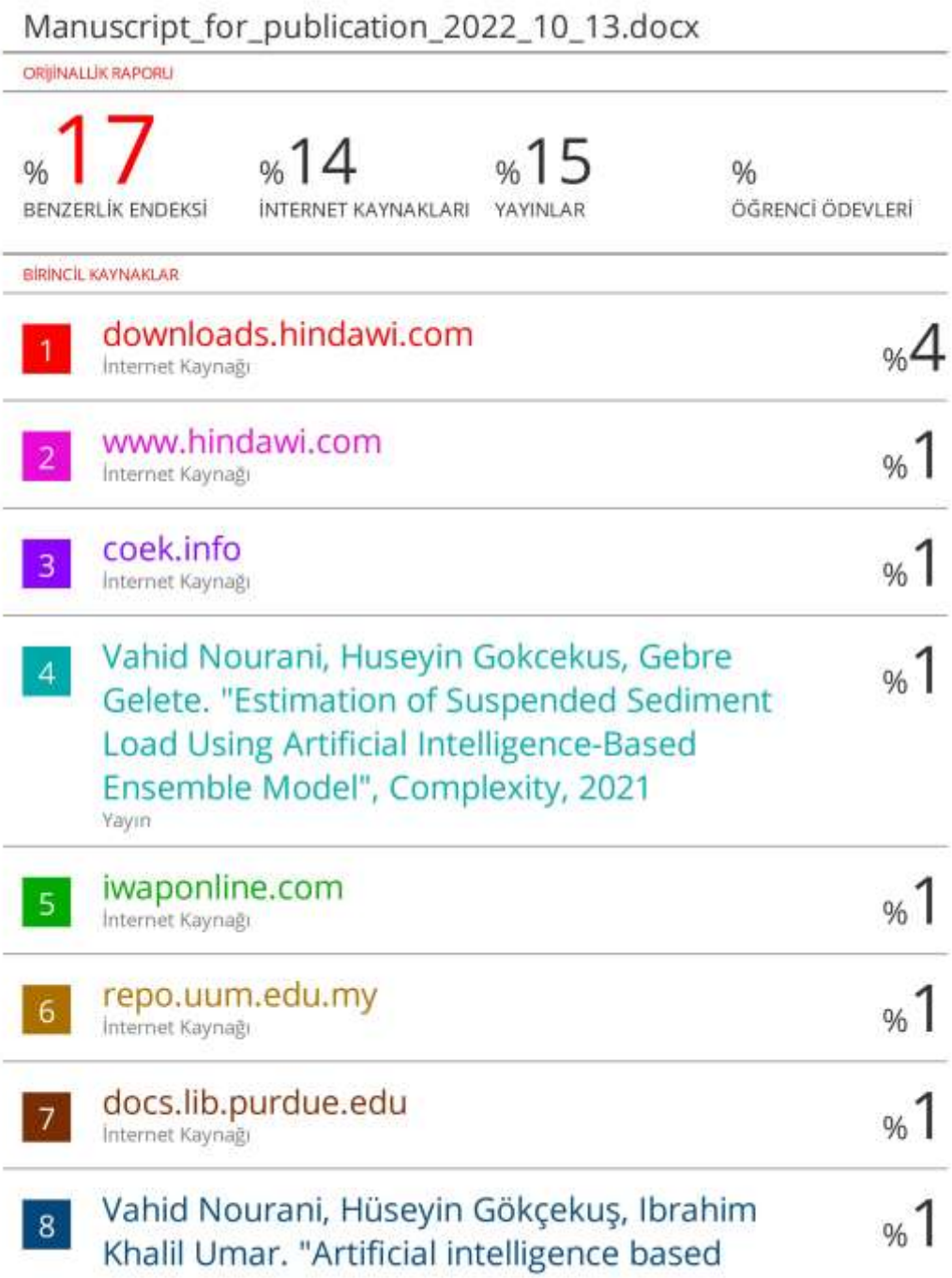
Model	MSE	RMSE	MAE	R2
kNN	10.022535211267606	3.165838784787944	2.1549295774647885	0.862830703
SVM	20.667761251166883	4.546180952312269	3.1132603510929093	0.717139205
Neural Network	10.080061251051733	3.174911219396809	2.344718404016604	0.862043396
AdaBoost (the Ensemble method)	4.222535211267606	2.0548808265365675	0.8084507042253521	0.94221001

Feature Statistics Tue Dec 20 22, 12

	Name	Distribution	Mean	Median	Dispersion	Min.	Max.
N	mask_use		9.28880	7.571	0.80552	0.00	31.714
N	all_bed_cap		596.84426	494.000	0.87973	0.00	2136.429
N	new_cases		600.66	473	0.90	0	2372
N	Pneumonia_st		37.963	12.6	2.310	3.9	724.0
N	new_deaths		9.50	8	0.90	0	47
N	Neural Network		9.35425	7.68206	0.819166	-0.56783	37.1607
N	AdaBoost (the Ensemble method)		9.47324	8	0.907314	0.00	47
N	SVM		8.46315	7.41153	0.732765	-0.134613	22.271
N	kNN		9.4338	7.5	0.849621	0.00	37.5
C	Fold			1	0		


Correlations
Tue Dec 20 22, 12:
Spearman correlation
1 +0.962 AdaBoost (the Ensemble method) new_deaths
2 +0.944 kNN new_deaths
3 +0.910 Neural Network new_deaths
4 +0.886 mask_use new_deaths
5 +0.886 SVM new_deaths
6 +0.815 all_bed_cap new_deaths
7 +0.796 new_cases new_deaths
8 -0.697 Pneumonia_st new_deaths

Appendix B: Turnitin Similarity Report



Manuscript_for_publication_2022_10_13.docx

ORİJİNALLİK RAPORU

% 17	% 14	% 15	%
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	downloads.hindawi.com İnternet Kaynağı	%4
2	www.hindawi.com İnternet Kaynağı	%1
3	coek.info İnternet Kaynağı	%1
4	Vahid Nourani, Huseyin Gokcekus, Gebre Gelete. "Estimation of Suspended Sediment Load Using Artificial Intelligence-Based Ensemble Model", Complexity, 2021 Yayın	%1
5	iwaponline.com İnternet Kaynağı	%1
6	repo.uum.edu.my İnternet Kaynağı	%1
7	docs.lib.purdue.edu İnternet Kaynağı	%1
8	Vahid Nourani, Hüseyin Gökçekuş, Ibrahim Khalil Umar. "Artificial intelligence based	%1

Appendix C: Curriculum Vitae (CV)

Curriculum Vitae (CV) **Long term career goal:** is to become among best data scientists in the world

Personal Details

Name: **Kedir Hussein Abegaz** (PhD, Asst. Prof)
Nationality: Ethiopian
Date of Birth: April 9, 1988
Sex: Male

Addis Ababa, Ethiopia
kedir6300@gmail.com
+251913012630/
+905338731645

International Profiles

		h-index	Citation
	ORCiD		
RG	Research Gate	19	8,635+
	Google Scholar	20	13,390+
	Web of Sciences	14	6,561+
SC	Scopus	16	4641+
AD	AD Scientific Index	Top 2% in Ethiopia	

Summary

Kedir as a researcher:

- I am an international academic editor and reviewer for over 10 journals,
- I have participated and published more than 50 researches and project works.
- I have successfully managed accomplished more than five projects on time and Budget.
- I spoke at more than 10 national and international conferences

Kedir as academician:

- I have completed all of my degrees (BSc, MSc, and PhD) with honors
- I lectured at more than ten universities and colleges in different modalities.
- I have mentored and assessed many more Masters Candidates.

Kedir on management:

- I am working as a Research Team Leader
- I have worked as a project coordinator and manager
- I have served as a Dean of post graduate programs, focal person of the school and department head positions

Work Experiences

1. **Senior Researcher and Team Lead; Frontier-*i* Consult,** Global office, Ethiopia (**Since Jan 2022**)

   Main Achievements:

   - I am currently leading a research team
   - Developed more than fifteen Technical proposals worth from $40 thousand to $450 thousands each for WBG, UNICEF, RTI, USAID, KOICA, JSI, FAO, UN Women, etc.
   - I have successfully coordinated and managed more than five projects for our clients
   - I deliver almost all projects with the approved budget and allocated time schedule.
   - I have provided proposal and report writing training for the staff members.
   - Lead so many problem solving and appreciation meetings

2. **Head of Post Graduate Programs; Yanet College**, Addis Ababa, Ethiopia (**June 2021 to now**)

   Main Achievements:

   - Worked as a postgraduate Dean.
   - Provided research software application for MSc/MPH students.
   - Advised and examined researches for more than 50 MSc/MPH graduated student.

4. **Research assistant and registry coordinator; American cancer society (AAU-CHS) (Jul 2018 to Dec 2019)**

Main Achievements:

- Lead the registry and coordinate the data collection for 1 year and half
- Developed more than 3 Technical proposals
- Engaged in routine data analysis and report writing for ACS-USA and MLU-Germany
- Developed 3 manuscripts for publication
- Assisted 2 batches of Oncology specialty students on thesis development.

5. **Research Consultant; Innovation and Technology Institute**, Ethiopia **(Nov 2018 to Feb 2019)**

Main Achievements:

- Advises and Reviewed 3 proposals for the Institute
- Provided a training on EndNote referencing and Stata
- Published 3 reports with the R&D staff members

6. **Assistant Professor and Instructor; Madda Walabu University**, Ethiopia **(Sep 2010 to Jan 2020)**

Main achievements:

- Lectured for more than ten years
- won 2 Research grants
- Train 3 times to academic staffs on EpiData, EndNote referencing and Stata
- Advised more than 20 MSc, and MPH students

Education

1. **PhD in Biostatistics**	**Near East University,** Northern Cyprus, Europe, Since January 2020 **(Artificial Intelligence and Machine learning speciality)**
2. **Assistant Professorship**	**Madda Walabu University**, Ethiopia, Since June 2018
3. **MSc in Biostatistics and Health Informatics**	**Mekelle University**, Ethiopia, March 2013 to December 2015 **(Data mining speciality)**
4. **BSc in Public Health**	**Kea-Med Medical College**, Ethiopia, Sep 2011 to Aug 2014
5. **BSc in Applied Statistics**	**Dilla University**, Ethiopia, November 2008 to July 2010 **(Minor Economics)**

Awards or Certificates

1. **Training of Trainer**	Python for Machine Learning and AI, Ohio State University, USA, 2022
2. **Trainer**	Research Software (Epi-Data, Stata, SPSS, EndNote), Yanet College, 2021.
3. **Cancer staging (TNM)**	International agency for research on cancer (IARC), France, Nov 2018
4. **Peer Reviewer**	Springer nature, UK, Dec 2017 and Elsevier, Netherlands, Feb 2018
5. **Active Reviewer**	PLOS One, San Francisco, USA, Aug 2018
6. **Research Preparation**	Elsevier publisher, Netherlands, Feb 2018
7. **Publication Process**	Elsevier publisher, Netherlands, Jan 2018
8. **Plagiarism and Ethics**	Elsevier publisher, Netherlands, Feb 2018
9. **Research Data Mg't**	Elsevier publisher, Netherlands, Feb 2018
10. **Research Evaluator**	Goba Referral Hospital, Madda Walabu University, Ethiopia, 2017
11. **Conference Coordinator**	Madda Walabu University; Bale Robe; Feb, 2017
12. **Research Assistant**	Leuphana University of Luneburg, World Bank, Ethiopia; Dec, 2016
13. **Higher Diploma Award**	Madda Walabu University; Ethiopia; July, 2016
14. **STATA and WEKA**	Ethio-Lens Institute of Technology, Mekelle; Dec 2013
15. **SPSS and ORANGE**	Ethio-Lens Institute of Technology, Mekelle; Dec 2013

Conferences

1. **Delegate**	Artificial Intelligence conference, ASTU, Adama, Jan27-29, 2022
2. **Panelist**	Yanet College, Addis Ababa, November 18-20, 2021
3. **Speaker**	University of Kyrenia, Northern Cyprus, June 23-26, 2021

4. **Speaker** DAAD summer course on NCD-Cancer, MLU-AAU, Sep 2019
5. **Collaborator** Global Burden Disease (GBD), USA, since 2019
6. **Panelist** GRH, Madda Walabu University, Ethiopia, March 2019
7. **Presenter** Ethiopian Statistical Association, Bahir-dar, Ethiopia; March 2018
8. **Presenter** Ethiopian Statistical Association, Addis Ababa, Ethiopia; Feb 2018
9. **Presenter** GRH, Madda Walabu University, Ethiopia, 2018
10. **Presenter** FBE, Madda Walabu University, Ethiopia, May 2018

Academia and research memberships

1. Guest Editor: Special issue on "*Biostatistics in Diagnostics Medicine*". **MDPI**, Switzerland, Since 2022. https://www.mdpi.com/journal/diagnostics/special_issues/G07EN1P21M
2. **Research Collaborator:** Global burden disease (**GBD Collaborator**), IHME, USA
3. **Editorial Board**; Journal of Big Data Research (JBR); Open access pub, USA; since Feb 2018 https://oap-cancer.org/journal/jbr/editorial-board
4. **Editorial Board**; OA Journal-Foods; USA, since June 2018 http://oa.enpress-publisher.com/index.php/FO/about/editorialTeam
5. **Editorial Board**; Probe Agricultural research, Singapore, since January 2018 http://www.front-sci.com/index.php/par/about/editorialTeam
6. **Valued Reviewer**; BMC Public Health; Springer Nature; UK, since January 2018
7. **Valued Reviewer**; BMC Women's Health; Springer Nature; UK, since January 2019
8. **Valued Reviewer**; BMC Agriculture and food security, Springer nature, UK, Since August 2018
9. **Valued Reviewer**; BMC Health services research, since December 2018
10. **Active Reviewer**; PLOS one; San Francisco; USA, since July 2018
11. **Regular member**; International statistical Institute (ISI); the Hague, Netherlands, since 2016
12. **Regular member**; Ethiopian Public Health Association (EPHA), Ethiopia, since 2017
13. **Regular Member**; Ethiopian cancer association (ECA), Ethiopia, since 2018

Regular member; Ethiopian Statistical Association (ESA); Ethiopia, since 2012

Training/Conferences

1. Theory and Practical uses of Neural Networks, Global One Health summer institute 2022, Ohio State University, USA.
2. The 2nd Deep learning Indaba-X Ethiopia on Artificial Intelligence, ASTU, Adama, Jan 27-29, 2022
3. The 6th international researchers, Statisticians and Young Statisticians congress, University of Kyrenia, Northern Cyprus, June 23-26, 2021
4. Mental Health Effects of the COVID-19 Pandemic, Harvard Medical School, June 11, 2020
5. **The 4th DAAD/PAGEL summer school, NCD and Cancer Epidemiology in Ethiopia: Enhancing Screening Approaches for Prevention, Elilly int. Hotel, Addis Ababa, Ethiopia**
6. **Eighth International Policy Conference on the African Child organized by the African Child Policy Forum(ACPF), UNCC, May 2019**
7. Ethiopian Statistical Association (ESA) conferences 25th, 26th, 27th and 28th from 2016-2019.
8. The 29th Ethiopian Public Health Association (EPHA) conferences, Feb 2018
9. Goba Referral Hospital of 2nd, 3rd and 4th Yearly national conferences from 2017-2019
10. College of business and Economics, 1st and 2nd yearly national conferences since 2017 and 2018
11. 1st international conference of statistics on agriculture (Haramaya University) (ICASA); Mar, 2016
12. Coping with Emerging Big Data: Bahir Dar University, Ethiopia, May 2017
13. TOT Training on "Peer Education on Reproductive Health and HIV" HAPCO of MWU, January 2016.
14. Statistics and Data Mining software: WEKA, Orange, Stata, SPSS, HMIS, Smart-Care, …from 2007-2017
15. The 1st international conference of statistics (UNECA and UNCH in Addis Ababa); May 20-23, 2016

Publications as PI (Participated in 55+ Researches since 2023)

1. Abegaz KH and Etikan I. *Boosting the Performance of Artificial Intelligence-Driven Models in Predicting COVID-19 Mortality in Ethiopia*. MDPI *Diagnostics* 2023, *13*, 658.

2. Abegaz KH and Etikan I. *Artificial Intelligence-Driven Ensemble model to predict mortality due to COVID-19 in east Africa.* MDPI *Diagnostics* 2022, *12*, 2861.
3. Abegaz KH. *Cancer incidence rates and trends in Addis Ababa, 2012-2016: Addis Ababa population-based cancer registry***.** International Journal of infectious disease (2021), 101 (S1) 219-264

4. Abegaz KH and Habtewold EM. *Trend and barriers of antenatal care utilization from 2000 to 2016 Ethiopian DHS: a data mining approach:* Elsevier, Scientific African (2019) 3:63
5. Abegaz KH. *Prevalence of Undernourishment: trend and contribution of East African countries to sub-Saharan Africa from 1991 to 2015***.** BMC: Agriculture and Food security (2018) 7:49
6. Abegaz KH and Mohammed AA. *Healthcare Expenditure and GDP in Ethiopia from 1995-2014: a Time Series Analysis.* BMC: Agriculture and & Food security (2018) 7:47
7. Abegaz KH. *Determinants of food security: evidence from Ethiopian rural household survey (ERHS) using pooled cross-sectional study*. BMC: Agriculture and & Food security (2017) 6:70
8. Welteji D, Mohammed K, Abegaz KH. *The contribution of productive safety net program for food Security of the rural households in the case of Bale Zone, Southeast Ethiopia.* BMC: Agriculture and & Food security (2017) 6:53
9. Abegaz KH, Berhe G, Semaw F. *Physicians' agreement in determining the causes of death using verbal autopsy data in Eastern Tigray, Ethiopia.* RJLBPCS (2017) 3:3
10. Abegaz KH, Seid MA, Jemal KM. *Food insecurity and agricultural shocks in the rural Ethiopia.* RJLBPCS (2017) 3:2
11. Abegaz KH, Atomssa EM. *Data mining of access to tetanus toxoid immunization among women of childbearing age in Ethiopia.* Machine Learning Research (2017) 2:2.
12. Abegaz KH, Habtewold EM. *The application of data-mining techniques on the utilization of antenatal care: proof from EDHS 2011.* BMC Public Health. In progress
13. Abegaz KH et.al. *Survival of Breast cancer patients at tikur anbessa specialized hospital, Ethiopia, Cox Regression model.* In progress

Personal competences

Languages	• Amharic (Mother tongue) • English (Fluency) • Tigrigna (Standard) • Arabic (Standard) • Afaan Oromo (Intermediate)
Interpersonal skills	• Able to work in a complex environment charged with multiple tasks, short deadlines and intense pressure to perform • Excellent analytical and problem-solving capacities with strong judgment • I have an excellent interpersonal, written and oral presentation skills A good team player accustomed to building team capacity and delegating working teams • Able to work independently and shoulder delegated authority • Able to communicate effectively, instilling trust and confidence
Computer skills	• Data entry software: **KOBO toolbox**, **Epi-Data** and **Epi-Info** • Statistical software: **STATA**, SPSS, and R • Data mining software: **Python**, Orange, Weka • Computer Graphics: CorelDraw… • Microsoft Office (Word, Excel, Access…)
Artistic skills	• First Dan degree in World Tae-Kwon-Do (WTF)

List of all published researches I participated

1. Boka A, Tadesse A, Hussein K. *Survival status and predictors of mortality among COVID-19 patients admitted to intensive care units at COVID-19 centers in Addis Ababa, Ethiopia: a retrospective study.* Annals of Medicine and Surgery. 2023 Jun;85(6):2368.
2. Feyisa JD, Woldegeorgis MA, Zingeta GT, Abegaz KH, Berhane Y. *Cervical Cancer Progression in Patients Waiting for Radiotherapy Treatment at a Referral Center in Ethiopia: A Longitudinal Study*. JCO Global Oncology. 2023 May;9:e2200435.
3. Abegaz, K.H. and İ. Etikan, *Boosting the Performance of Artificial Intelligence-Driven Models in Predicting COVID-19 Mortality in Ethiopia.* Diagnostics, 2023. **13**(4): p. 658.
4. GBD Collaborators., *Global investments in pandemic preparedness and COVID-19: development assistance and domestic spending on health between 1990 and 2026.* The Lancet Global Health, 2023. **11**(3): p. e385-e413.
5. GBD Collaborators., *The global burden of cancer attributable to risk factors, 2010–19: a systematic analysis for the Global Burden of Disease Study 2019.* The Lancet, 2022. **400**(10352): p. 563-591.
6. GBD Collaborators., *Progress in health among regions of Ethiopia, 1990–2019: a subnational country analysis for the Global Burden of Disease Study 2019.* The Lancet, 2022. **399**(10332): p. 1322-1335.
7. GBD Collaborators., *The burden of injury in Central, Eastern, and Western European sub-region: a systematic analysis from the Global Burden of Disease 2019 Study.* Archives of public health, 2022. **80**(1): p. 1-14.
8. Abegaz, K.H. and İ. Etikan, *Artificial Intelligence-Driven Ensemble Model for Predicting Mortality Due to COVID-19 in East Africa.* Diagnostics, 2022. **12**(11): p. 2861.
9. GBD Collaborators., *Tracking development assistance for health and for COVID-19: a review of development assistance, government, out-of-pocket, and other private spending on health for 204 countries and territories, 1990–2050.* The Lancet, 2021. **398**(10308): p. 1317-1343.
10. GBD Collaborators., *Mapping routine measles vaccination in low-and middle-income countries.* Nature, 2021. **589**(7842): p. 415-419.

11. GBD Collaborators., *Mapping geographical inequalities in oral rehydration therapy coverage in low-income and middle-income countries, 2000–17.* The Lancet Global Health, 2020. **8**(8): p. e1038-e1060.
12. GBD Collaborators., *Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019.* The Lancet, 2020. **396**(10258): p. 1160-1203.
13. GBD Collaborators., *Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019.* The Lancet, 2020. **396**(10258): p. 1204-1222.
14. GBD Collaborators., *Mapping geographical inequalities in childhood diarrhoeal morbidity and mortality in low-income and middle-income countries, 2000–17: analysis for the Global Burden of Disease Study 2017.* The Lancet, 2020. **395**(10239): p. 1779-1801.
15. GBD Collaborators., *Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019.* The lancet, 2020. **396**(10258): p. 1223-1249.
16. GBD Collaborators., *Five insights from the global burden of disease study 2019.* The Lancet, 2020. **396**(10258): p. 1135-1159.
17. Megersa, S. and Abegaz KH et al., *The Association Between Diarrheal Morbidity and ODF Status Among Under Five Children in Southeast, Ethiopia.* 2020.
18. GBD Collaborators, *Author Correction: Mapping local patterns of childhood overweight and wasting in low-and middle-income countries between 2000 and 2017.* 2020.
19. GBD Collaborators., *Estimating global injuries morbidity and mortality: methods and data used in the Global Burden of Disease 2017 study.* Injury Prevention, 2020. **26**(1): p. l125-l153.
20. GBD Collaborators., *Mapping geographical inequalities in access to drinkingwater and sanitation facilities in low-income andmiddle-income countries, 2000–17.* 2020.
21. GBD Collaborators., *Mapping local patterns of childhood overweight and wasting in low-and middle-income countries between 2000 and 2017.* 2020.

22. GBD Collaborators., *Estimating global injuries morbidity and mortality: methods and data used in the Global Burden of Disease 2017 study.* Injury Prevention, 2020. **26**(Suppl 2): p. i125-i153.
23. GBD Collaborators., Diarrhoea, L.B.D., et al., *Mapping geographical inequalities in childhood diarrhoeal morbidity and mortality in low-income and middle-income countries, 2000-17: analysis for the Global Burden of Disease Study 2017.* Lancet, 2020.
24. GBD Collaborators., *Mapping geographical inequalities in access to drinking water and sanitation facilities in low-income and middle-income countries, 2000–17.* The Lancet Global Health, 2020. **8**(9): p. e1162-e1185.
25. GBD Collaborators., *The global distribution of lymphatic filariasis, 2000–18: a geospatial analysis.* The Lancet Global Health, 2020. **8**(9): p. e1186-e1194.
26. Collaborators, L.D.B.o.M., *Author Correction: Mapping local patterns of childhood overweight and wasting in low-and middle-income countries between 2000 and 2017.* Nature medicine, 2020. **26**(8): p. 1308.
27. Collaborators, L.B.o.D.W., *Mapping geographical inequalities in access to drinking water and sanitation facilities in low-income and middle-income countries, 2000–17.* The Lancet. Global Health, 2020. **8**(9): p. e1162.
28. Collaborators, L.B.o.D.D., *Department of Error: Mapping geographical inequalities in childhood diarrhoeal morbidity and mortality in low-income and middle-income countries, 2000–17: analysis for the Global Burden of Disease Study 2017 (The Lancet (2020) 395 (10239)(1779–1801),(S0140673620301148),(10.1016/S0140-6736 (20) 30114-8)).* The Lancet, 2020. **395**(10239): p. 1762.
29. Collaborators, G. and J. Ärnlöv, *Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019.* The Lancet, 2020. **396**(10258): p. 1160-1203.
30. GBD Collaborators., *Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019.* 2020.
31. Al-Aly, Z. and L.B.o.D.D. Collaborators, *Mapping geographical inequalities in oral rehydration therapy coverage in low-income and middle-income countries, 2000-17.* 2020.

32. Abegaz, K., *Cancer incidence rates and trends in Addis Ababa, 2012–2016: Addis Ababa population-based cancer registry.* International Journal of Infectious Diseases, 2020. **101**: p. 248.
33. GBD Collaborators *Mapping local patterns of childhood overweight and wasting in low- and middle-income countries between 2000 and 2017.* Nature medicine, 2020. **26**(5): p. 750-759.
34. GBD Collaborators., *The global, regional, and national burden of colorectal cancer and its attributable risk factors in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017.* The Lancet: Gastroenterology and Heptology, 2019. **4**(12).
35. GBD Collaborators., *The global, regional, and national burden of pancreatic cancer and its attributable risk factors in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017.* The lancet Gastroenterology & hepatology, 2019. **4**(12): p. 934-947.
36. GBD Collaborators., *Estimating global injuries morbidity and mortality: methods and data used in the Global Burden of Disease 2017 study.* journal online injuryprev, 2019.
37. GBD Collaborators., *The global, regional, and national burden of colorectal cancer and its attributable risk factors in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017.* https://doi. org/10.1016/S2468-1253 (19) 30345-0, 2019.
38. Abegaz, K.H. and E.M. Habtewold, *Trend and barriers of antenatal care utilization from 2000 to 2016 Ethiopian DHS: a data mining approach.* Scientific African, 2019. **3**: p. e00063.
39. Abegaz, K.H. and A.A. Mohammed, *Healthcare expenditure and GDP in Ethiopia from 1995 to 2014: a time-series analysis.* Agriculture & Food Security, 2018. **7**: p. 1-7.
40. Abegaz, K.H., *Exploring trend and barriers of antenatal care utilization using data mining: evidence from EDHS of 2000 to 2016.* bioRxiv, 2018: p. 351858.
41. Abegaz, K.H., *Prevalence of undernourishment: trend and contribution of East African countries to sub-Saharan Africa from 1991 to 2015.* Agriculture & Food Security, 2018. **7**: p. 1-6.

42. Diriba Welteji, K.M. and K.H. Abegaz, *The contribution of Productive Safety Net Program for food security of the rural households in the case of Bale Zone, Southeast Ethiopia.* Agriculture & Food Security, 2017. **6**(53).
43. Abegaz, K.H., G. Berhe, and S. Ferede, *Physicians' agreement in determining the causes of death using verbal autopsy data in Eastern Tigray, Ethiopia.* Res J Life Sci Bioinformatics, Pharm. Chem Sci (Camb), 2017. **3**: p. 100-14.
44. Abegaz, K.H. and E.M. Atomssa, *Data mining of access to tetanus toxoid immunization among women of childbearing age in Ethiopia.* Machine Learning Research, 2017. **2**(2): p. 54-60.
45. Abegaz, K.H., M. Adem Seid, and K.M. Jemal, *Food Insecurity and Agricultural Shocks in The Rural Ethiopia*. 2017.
46. Abegaz, K.H., *Determinants of food security: evidence from Ethiopian Rural Household Survey (ERHS) using pooled cross-sectional study. Agric & Food Secur. 2017; 6: 70*. 2017.
47. Collaborators., *Global burden of 369 diseases and injuries in 204 countries and territories, 1990 2019: a systematic analysis for the Global Burden of Disease Study 2019 The Lancet, 2020.* The lancet. **396**(10258): p. 1204-1222.

List of Technical proposals I developed from 2020-2023

1. An assessment on post project effectiveness of Productive Safety Net Program (PSNP) participants' linkage with financial service providers and sustainability of Private Service Providers (PSPs), **WBG**
2. Baseline Data Collection of the Impact Evaluation of the UPSNJP Bikat Program, **WBG**
3. Baseline survey for rural agricultural value-chain improvement project through linking smallholder farmers with rural transformation center (RTC), east Gojam zone, Amhara regional state, Ethiopia, **KOIKA**
4. Baseline Survey on Gender in Emergencies (GiE) in South Sudan program Wau County, Western Bahr El Ghazel State Unity State**, CARE South Sudan**
5. Collect Quantitative Data from Randomly Selected 500 Households from 5 Woredas (districts) and 20 kebeles in the Awash basin, Ethiopia" **IWMI**
6. Conducting Advanced COVID-19 Vaccine Data Analytics, **JSI**
7. Conducting Gender Analysis in PReSERVE Project Woredas of Amhara Region, **RTI**
8. Conducting School Visits and Baseline Household Survey in Rural Amhara and Oromia Regions of Ethiopia, **IFPRI**
9. Data Collection in Sudan on Migrants along the Northern Corridor under the EU-IoM Joint Initiative Programme In the HoA Region, **IOM**
10. Designing nutrition-sensitive Cash Plus package for smallholder farmers (including PSNP clients) to improve nutritive home consumption and supply HGSF programme, **FAO**
11. EatSafe Ethiopia Baseline Assessment in Hawassa technical proposal, **GAIN**

12. End-line survey to assess the impact and cost effectiveness of different content delivery approaches on technology adoption, **IFPRI**
13. Food and Agricultural System Mapping and Analysis (FASMA) for Lot 1- Dire Dawa and Bahir Dar, **RTI**
14. Geared for Success Baseline Assessment, **Oxfam Canada and WCC**
15. Gender Assessment on a project "Horticulture for Growth (H4G): an initiative of the agricultural commercialization clusters (ACCs), **TechnoServe**
16. Gender Audit in GWPTC Students, Trainers and Administrative Staff, **EASTRIP WBG**
17. Identifying and assessing the socio-economic characteristics of migrant, refugee, and undocumented populations in the Republic of Djibouti, **WBG**
18. Implementation of a mini Household survey using LQAS (HPF3), **Grown Agents**
19. Listing and Baseline Survey for Land Rental Market Impact Evaluation Study, **WBG**
20. Midline review of the NCA South Sudan Country Strategy 2020-2024 , **Norwegian Church**
21. Monitoring Coverage for Mass Drug Administration in South Sudan, **CBM**
22. Primary data collection for the tobacco control data initiative in regional states of Ethiopia, **Development gateway**
23. Primary quantitative data collection in Amhara region, **PreSERVE, RTI**
24. Providing Survey Management Services in Support of a National Population-Based HIV Impact Assessment-Ethiopia, **ICAP Columbia University**
25. Research firm to implement a Baseline Survey for the USAID Healthy Behaviors Activity in Ethiopia, **FHI 360-2022**
26. Survey & Investigation of identified villages in Ethiopia for development of Solar Mini-grid projects and preparation of Bankable DPRs for these sites, **TERI, ISA**
27. Terminal Evaluation for the Youth Action for Reduced Violence and Enhanced Social Cohesion in Wau, South Sudan, **IoM and UNESCO south Sudan**
28. The causes, trends, effects, and impacts of migration in Addis Ababa city administration, **Addis Ababa city administration**
29. An endline survey for Digital agricultural services (DAAS) for IFPRI project: 2023
30. Evaluation of Electronic Medical Records (EMR) System Usage and User satisfaction at Selected Health Facilities in Four Regions of Ethiopia, **ICAP**

List of projects I managed (from 2020 to date)

1. Baseline Data Collection of the Impact Evaluation of the UPSNJP Bikat Program, **WBG**
2. Conducting Baseline and Follow-up Surveys for the PWs, PDS, and Livelihood Impact evaluation, **WBG**
3. Designing nutrition-sensitive Cash Plus package for smallholder farmers (including PSNP clients) to improve nutritive home consumption and supply HGSF programme, **FAO**
4. DRDIP-I End of project evaluation and Baseline survey for DRDIP-II, **MoA**
5. Smallholder Farmer Endline Survey Data Collection for Soufflet-IFC Malt Barley Value Chain Advisory Project in Ethiopia, **IFC-2023**
6. Conduct Health Facility Baseline Survey on Prevention and Management of Health Complication of Female Genital Mutilation (FGM)**, UN Children Fund**

7. Baseline Survey for the National Campaign for Promoting Knowledge, Attitude, and Behavioural Change in Population and Reproductive Health in Ethiopia (SHaPE Phase 2) for Target Groups in Two Cities and Five Regions, **KOICA**
8. Gender Audit for Women's Empowerment Through Gender Transformative Opportunities (WE-GO) Project, **CARE**
9. Estimating the Economic Costs of Intimate Partners' Violence (IPV) against women in Ethiopia**, UN Women**
10. Validating the Global Framework to Measure and Evaluate Changes in Social Norms in FGM Programming**, UNICEF**
11. Developing a Civil Registration Vital Statistic in South Sudan, **UNDP**
12. Food and Agricultural System Mapping and Analysis (FASMA) for Lot 1- Dire Dawa and Bahir Dar**, RTI**
13. Baseline Survey of Impact of Vegetable Seed and Complementary Training Provided to Households Affected by Drought in Ethiopia, **WorldVeg** and **CRS**
14. Consultancy services to conduct baseline survey, skill gap assessment and organization of workshops, **MoLS-Ethiopia**
15. Conducting a baseline study of civic engagement activities in Ethiopia, **Creative associates-USAID**
16. Impact and process evaluation of the Global Alliance for Improved Nutrition (GAIN) *Better Dairy for All* program in Ethiopia, **RTI** and **GAIN**
17. Strengthening federal and regional levels monitoring, evaluation and learning (MEL) system for the prevention of harmful practices and violence against children (VAC) in Ethiopia, **UNICEF** and **UNFPA**