12-19-2023

Advanced Deep Learning Multivariate Multi-Time Series Framework for a Novel COVID-19 Dataset

SWASTIK BAGGA
University of Windsor

Advanced Deep Learning Multivariate Multi-Time Series Framework for a Novel COVID-19 Dataset

By

Swastik Bagga

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2023

Advanced Deep Learning Multivariate Multi-Time Series Framework for a Novel COVID-19 Dataset

by

Swastik Bagga

APPROVED BY:

A. Hamdi Sakr
Department of Electrical and Computer Engineering

S. Samet
School of Computer Science

Z. Kobti, Advisor
School of Computer Science

December 06, 2023

DECLARATION OF CO-AUTHORSHIP / PREVIOUS PUBLICATION

1. Co-Authorship

   I hereby declare that this thesis incorporates material that is the result of research conducted under the supervision of Dr. Ziad Kobti. In all cases, the key ideas, primary contributions, experimental designs, data analysis, and interpretation were performed by the author, and the contribution of co-authors was primarily through the proofreading of the published manuscripts. The initial idea and inspiration for my thesis was generously contributed by Shaon Bhatta Shuvo.

   I am aware of the University of Windsor Senate Policy on Authorship, and I certify that I have properly acknowledged the contribution of other researchers to my thesis and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

   I certify that, with the above qualification, this thesis, and the research to which it refers, is a product of my own work.

2. Previous Publication

   This thesis includes the paper that has been published in a peer-reviewed conference, as follows:

Publication Title/Full Citation	Publication Status
Shuvo, S. B., Bagga, S., and Kobti, Z. (2023). COVID-19 Analysis in Canada using Deep Learning and Multi-Factor Data-Driven Approach with a Novel Dataset. At 2023 IEEE Symposium on Computers and Communications (ISCC).	Published

   I certify that I have obtained written permission from the copyright owner(s) to include the above-published material(s) in my thesis. I certify that the above

material describes work completed during my registration as a graduate student at the University of Windsor.

3. General

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained written permission from the copyright owner(s) to include such material(s) in my thesis. I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

This thesis introduces an innovative framework aimed at addressing the complexities of predicting outcomes in multivariate multi time series datasets in regression analysis. By applying this framework to a novel COVID-19 dataset, it enhances predictive analytics by providing accurate forecasts for epidemic trends at regional or provincial levels, going beyond national-level analysis. The framework incorporates advanced data preprocessing, feature selection, engineering, encoding, and model architecture, effectively capturing intricate variable interactions and temporal dependencies. This makes it a powerful tool for tackling multivariate multi time series regression challenges, offering valuable insights for informed decision-making.

Predicting outcomes in such datasets is challenging due to variable interconnections and temporal dynamics. The framework presented in the thesis adeptly models dependencies and latent patterns while considering real-world uncertainties. It demonstrates its practical value in localized epidemic trend forecasting, where deep data understanding is crucial for effective decision-making. Extensive experimentation shows that the framework outperforms traditional regression models and time series models in terms of various performance metrics, such as R^2, MAE, MaxAE, and RMSE. A novel model, DeepAREstimator, is introduced to balance performance and training time, offering a maintainable and scalable solution for real-world applications. The findings contribute to advancing predictive analytics, and providing essential insights for decision-making, particularly in localized epidemic trend forecasting.

DEDICATION

I would like to dedicate this thesis to my mom for her incredible love and support. Because I believe that she is the real backbone of our family, this is to appreciate her selfless hard work and efforts towards the family.

Furthermore, I dedicate it to my dad to raise me like a son and give me wings to fly. To my grandparents, for always trusting me and supporting me in my hard times, without their encouragement, nothing would have been easy. And to my entire family for their unconditional affection towards me.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my esteemed advisor, Dr. Ziad Kobti, whose remarkable research expertise and unwavering mentorship have had a profound impact on my academic journey. Dr. Kobti generously provided me with invaluable opportunities to delve into my research field, consistently offering encouragement and expert guidance that significantly enhanced the quality of my work. Their commitment to nurturing my growth, both as a researcher and as an individual, has been exceptionally inspiring. The constructive feedback they shared played a central role in honing my thesis and instilling within me a profound sense of self-assurance and achievement. The privilege of collaborating with Dr. Kobti has been an enriching experience, and I am profoundly thankful for the priceless opportunities and support I received throughout my academic pursuit. I would also like to extend my gratitude to Shaon Shuvo , whose initial motivation and ideas for the thesis served as a catalyst for this endeavor. Collaborating with Dr. Kobti and being inspired by Shaon Shuvo's insights has been an enriching experience, and I am profoundly thankful for the priceless opportunities and support I received throughout my academic pursuit.

Secondly, I also extend my sincere appreciation to my thesis committee members, Dr. Ahmed Hamdi Sakr and Dr. Saeed Samet, whose invaluable insights and inspiration played a significant role in the successful completion of this thesis.

My association with the University of Windsor has been a great honor. Moreover, I am thankful for the guidance offered by other esteemed professors at UWindsor during the course of my MSc program.

My heartfelt thanks go to Mrs. Monique Ritz, Mrs. Melissa Robinet, and Mrs. Christine Weisener for their unwavering support and invaluable assistance in resolving various academic matters.

Lastly, I owe a profound debt of gratitude to my family and friends. Also, I humbly extend my thanks to the School of Computer Science and all concerned people who helped me in this regard.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
DT	Decision Tree
KNN	K Nearest Neighbors
NB	Naive Bayes
LR	Logistic Regression
RF	Random Forest
MLP	Multi Layer Perceptron
LGB	Light Gradient Boosting Machines
NN	Neural Network
MAE	Mean Absolute Error
MaxAE	Max Absolute Error
RMSE	Root Mean Square Error
R^2	Coefficient of Determination
HCMVMT	High Cardinality Multivariate Multi-time series
ARIMA	AutoRegressive Integrated Moving Average
VAR	Vector Autoregression
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
SGD	Stochastic Gradient Descent
Adam	Adaptive Moment Estimation

RMSprop	Root Mean Square Propagation
Adagrad	Adaptive Gradient Algorithm
Momentum	Momentum Optimization
CNN	Convolutional Neural Network
SGD	Stochastic Gradient Descent
ReLU	Rectified Linear Unit
SGD	Stochastic Gradient Descent
Adam	Adaptive Moment Estimation
RMSprop	Root Mean Square Propagation
Adagrad	Adaptive Gradient Algorithm
Momentum	Momentum Optimization
DNN	Deep Neural Network
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning

CHAPTER 1

Introduction

1.1 Background

In recent years, the field of data science and predictive analytics has been transformed by a burgeoning wealth of multivariate multi time series datasets. These datasets are characterized by their complexity, involving numerous variables and intricate temporal dependencies. The intersection of these data characteristics poses substantial challenges for traditional regression prediction models. The conventional models that have been effective in simpler, univariate time series data struggles to cope with the intricate relationships, temporal dynamics, and intertwined nature of variables found in these complex datasets.

The relevance of this challenge extends across a spectrum of domains, including finance, healthcare, climate science, epidemiology, and more. In these fields, data is collected over time, with each data point being associated with a multitude of variables. For instance, consider healthcare, where patient data involves a myriad of physiological parameters, such as heart rate, blood pressure, and temperature, recorded at regular intervals. In epidemiology, understanding the dynamics of disease outbreaks requires analyzing data across regions, each characterized by a diverse set of factors, including population density, healthcare infrastructure, and local policies. In finance, stock prices, influenced by various external factors, illustrate the complex relationships between variables over time. Accurate prediction and understanding of outcomes in these datasets are imperative for making informed decisions and crafting effective strategies.

The intricacy of these datasets arises from the intricate web of dependencies between variables, both within the same time point and across different time points. Variables often exhibit non-linear and dynamic interactions, compounding the challenge. For example, in a financial dataset, the stock prices of different companies may be interrelated, with external economic events exerting further influence on these relationships.

Adding to this complexity are real-world uncertainties, which introduce an additional layer of challenges. Missing data, measurement errors, and unexpected events can introduce noise and disrupt regular patterns within the data. As such, it is imperative that any effective predictive model accounts for these uncertainties.

As a result, there is a growing demand for innovative approaches that can effectively handle multivariate multi time series data. Such approaches need to capture dependencies, patterns, and uncertainties while providing accurate predictions. The need for such models was prominently highlighted during the COVID-19 pandemic, where understanding and forecasting regional epidemic trends became a critical task. National-level analysis, while informative at a high level, does not provide the granularity required for localized decision-making. This underscores the importance of bridging the gap between national and regional-level analysis, which is one of the key motivations behind this research.

The importance of addressing this challenge is underscored by its practical applications. As we have seen during the COVID-19 pandemic, understanding and forecasting regional epidemic trends became an urgent and critical task. National-level analysis, while informative at a high level, could not provide the granularity required for localized decision-making. Thus, the need for models that can bridge the gap between national and regional-level analysis becomes evident.

Solving this problem has profound implications for informed decision-making and strategy development across diverse domains. It not only advances the field of predictive analytics but also empowers policymakers and decision-makers with the tools to make targeted and effective decisions. This research aims to develop a framework that is poised to meet this challenge, offering innovative solutions to enhance our

understanding and prediction of outcomes in multivariate multi time series datasets. The applications of this work span domains as varied as healthcare, finance, and epidemiology, underlining its critical importance in the current data-driven era.

In summary, the background of this thesis is rooted in the evolving landscape of data science, where complex multivariate multi time series datasets pose significant challenges to traditional regression prediction models. Understanding and effectively predicting outcomes in such datasets is essential for informed decision-making in various domains, with the COVID-19 pandemic serving as a prime example of the need for more localized and accurate forecasts. This research aims to develop a framework that addresses these challenges and provides valuable contributions to the field of predictive analytics.

1.1.1 Regression Prediction Problem

In the realm of predictive analytics, the Regression Prediction Problem stands as a fundamental and enduring challenge. At its core, it is a task concerned with understanding and modeling the relationships between a dependent variable and one or more independent variables. In essence, this problem seeks to answer the question: "Given a set of input variables, what can we predict about the outcome?" This outcome, typically a continuous numerical value, may represent various phenomena, such as stock prices, patient health indicators, or economic indicators.

Traditional regression models, like linear regression, are often employed to address this problem, assuming that the relationships between variables are linear. However, as datasets have become increasingly complex, with a multitude of variables and intricate temporal dependencies, the linear assumptions underlying traditional regression models are frequently challenged. Consequently, there arises a need for innovative approaches, particularly in the context of multivariate multi time series datasets.

1.1.2 Regression Time Series Prediction Problem

Extending the scope of the Regression Prediction Problem, the Regression Time Series Prediction Problem addresses scenarios where data is collected over time. This problem encapsulates an array of challenges distinct from standard regression due to the temporal nature of the data. Specifically, it involves modeling how a dependent variable evolves over time, given one or more independent variables.

The significance of the Regression Time Series Prediction Problem becomes all the more apparent in the context of multivariate multi time series datasets. In these datasets, variables exhibit temporal dependencies, and the interactions among them can be highly dynamic and nonlinear. For instance, in the case of forecasting regional COVID-19 trends, the number of cases in one region may influence the number of cases in neighboring regions over time. Additionally, external factors, like vaccination campaigns or policy changes, may introduce non-trivial temporal dependencies into the dataset.

As the world grapples with increasingly complex data, characterized by intricate relationships between variables and intricate temporal dynamics, the need for novel solutions to the Regression Time Series Prediction Problem intensifies. These solutions must address the challenges of modeling dependencies, handling real-world uncertainties, and providing accurate predictions, all while considering the implications for informed decision-making.

Solving the Regression Time Series Prediction Problem is of paramount importance in various domains, including epidemiology, finance, and healthcare. Effective solutions are not only poised to advance the field of predictive analytics but also to empower decision-makers with the tools to make targeted and effective decisions, particularly in scenarios that necessitate localized epidemic trend forecasts, regional financial predictions, and patient health monitoring. This research endeavors to provide innovative and robust solutions to this challenge, shaping the landscape of predictive analytics in the face of evolving data complexities.

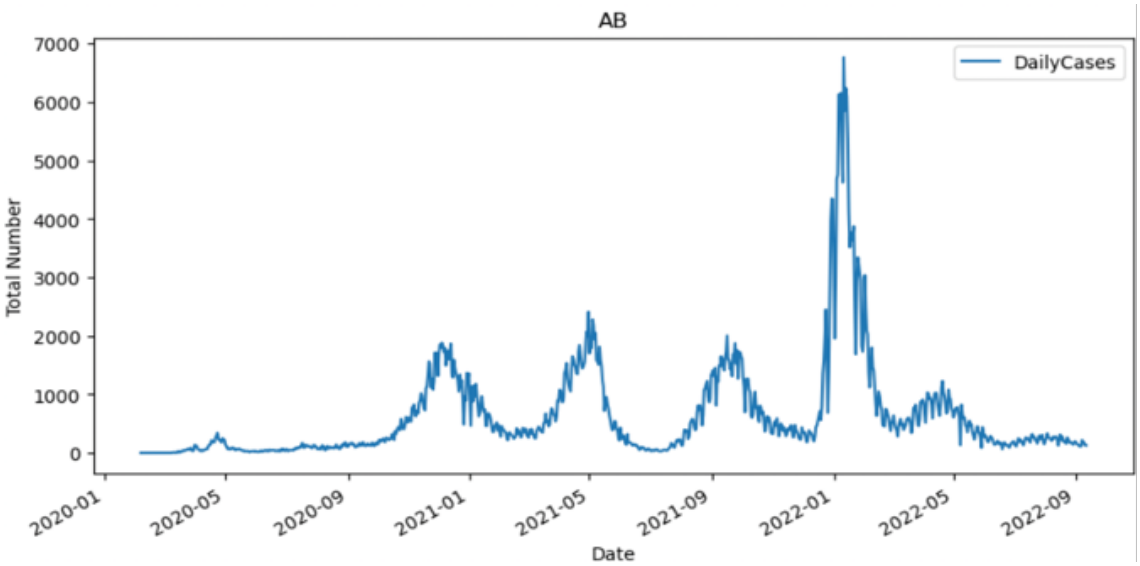


Fig. 1.1.1: Number Of COVID-19 Daily Cases in Alberta Time Series

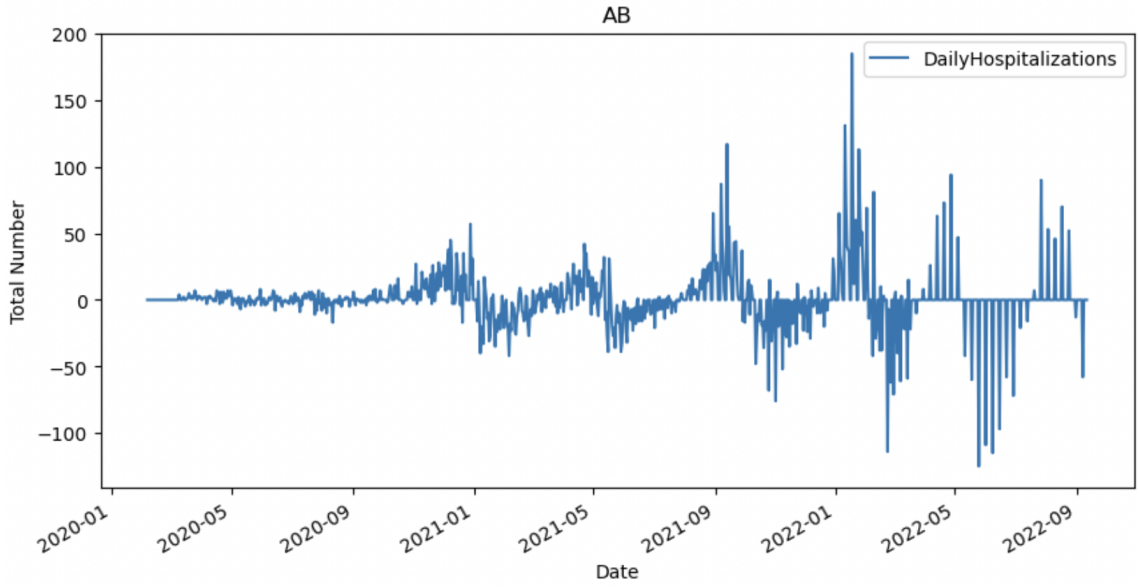


Fig. 1.1.2: Number Of COVID-19 Daily Hospitalizations in Alberta Time Series

1.1.3 State-of-the-Art Prediction Algorithms for Regression Time Series

In the domain of regression time series prediction, several advanced algorithms and models have risen to prominence, each offering unique capabilities and suitability for different scenarios. Here, we provide more in-depth descriptions of these state-of-the-art techniques:

1.1.3.1 Deep Neural Networks (DNNs)

Deep Neural Networks represent a class of machine learning models characterized by their depth and capacity to learn intricate, non-linear relationships within data. These networks consist of multiple hidden layers, enabling them to capture complex patterns and dependencies in regression time series data. DNNs[30] have demonstrated remarkable adaptability, proving effective in a wide range of domains. Their ability to handle high-dimensional data and learn hierarchical features makes them invaluable for modeling intricate temporal dynamics.

1.1.3.2 Recurrent Neural Networks (RNNs)

Recurrent Neural Networks are specialized deep learning models designed for sequential data, making them well-suited for time series analysis. What sets RNNs[23] apart is their capacity to maintain a memory of past time steps. This memory mechanism enables them to capture temporal dependencies within the data, which is essential for accurate regression time series predictions. Variants such as **Long Short-Term Memory (LSTM)[20]** and **Gated Recurrent Unit (GRU)[14]** have been introduced to address issues like vanishing gradients, providing more robust solutions for modeling longer-range dependencies.

1.1.3.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a non-parametric and instance-based algorithm. It operates on the principle of similarity, where predictions are made by identifying the

k-nearest data points to a given instance and using their values to infer the target value. KNN[22] is particularly effective when dealing with local patterns and small-scale trends in regression time series data. It is non-parametric, meaning it doesn't make strong assumptions about the data distribution, allowing it to adapt to various patterns. However, its effectiveness may diminish when facing data with intricate dependencies and substantial noise.

1.1.3.4 ARIMA/VAR Models

AutoRegressive Integrated Moving Average (ARIMA)[19] and **Vector Autoregression (VAR)[24]** models are classical approaches with a strong foundation in time series analysis. ARIMA[19] models are well-suited for capturing linear temporal dependencies within univariate time series data. They consist of auto-regressive, integrated, and moving average components. VAR models extend this capability to multivariate time series data, making them valuable for studying the interactions between multiple variables over time. These models are effective in situations where the data exhibits clear and stationary patterns, making them valuable for simple, linear relationships. However, they may struggle when dealing with more complex, non-linear relationships and non-stationary data.

1.1.3.5 Prophet

Prophet[36], developed by Facebook, is a time series forecasting model designed for datasets with daily observations and seasonal patterns. It is characterized by its ability to handle data with missing values, outliers, and holidays effectively. This makes it a valuable tool in situations where data is noisy or has gaps. Prophet has found applications in various domains, including e-commerce, social media analytics, and more. Its adaptability to handling real-world complexities and noisy data makes it a noteworthy addition to the suite of time series forecasting tools.

Each of these state-of-the-art prediction algorithms for regression time series offers unique strengths and capabilities. The choice of algorithm depends on the specific characteristics of the data, the nature of the relationships between variables, and the

objectives of the analysis. This thesis will investigate the performance and limitations of these algorithms within the context of multivariate multi time series datasets, with a focus on localized epidemic trend forecasts. The research aims to harness the advantages of these state-of-the-art techniques while addressing the complexities and intricacies of the data at hand, thereby advancing the field of predictive analytics.

1.1.4 High Cardinality Multivariate Multi-time series Datasets

The advent of the data-driven era has ushered in a deluge of information across diverse domains, and the intricacy of this data often defies simple categorization. Among the most complex and challenging datasets encountered are those characterized by high cardinality, multivariate attributes, and temporal dependencies. These datasets, often referred to as High Cardinality Multivariate Multi-time series datasets, pose a unique set of challenges and opportunities for data analysts and machine learning practitioners.

1.1.4.1 Complexity and Challenges

These datasets present a unique set of complexities and challenges that have direct implications for the regression prediction problem and, more specifically, the prediction of epidemic trends at regional or provincial levels, as explored in this thesis. The complexity arises from several key factors:

- **Interwoven Variables:** The term "high cardinality" in this context refers to the presence of a vast number of unique entities, such as regions or provinces. Each entity is associated with a multitude of variables. For instance, in the context of COVID-19 prediction, each region might be characterized by data on population density, healthcare infrastructure, local policies, and more. The interplay between these variables is far from linear, making it a challenge to understand their relationships

- **Temporal Dynamics:** High cardinality multivariate multi-time series datasets involve data collected over time, and each time point is associated with a multi-

tude of variables. In the case of the COVID-19 dataset, this temporal aspect is particularly relevant, as the number of cases, hospitalizations, and testing rates evolve over time. This temporal aspect introduces dynamic dependencies and patterns that require sophisticated modeling techniques.

- **Real-World Uncertainties:** In these datasets, uncertainties often abound. Data can be missing, noisy, or subject to unforeseen events. For instance, disruptions in testing capacity, changes in reporting standards, or regional lockdowns during the COVID-19 pandemic introduce uncertainty into the dataset. Addressing these real-world uncertainties is paramount for building reliable prediction models.

- **Non-Linear Dependencies:** High cardinality multivariate multi-time series datasets often exhibit complex non-linear dependencies between variables. Traditional linear regression models may not effectively capture these intricate relationships. Understanding and modeling these non-linear interactions pose a significant challenge.

- **High Dimensionality:** With a multitude of variables for each entity and multiple time points, these datasets are characterized by high dimensionality. High-dimensional data introduces computational challenges and may require dimensionality reduction techniques to avoid the curse of dimensionality.

- **Heterogeneity of Entities:** Each entity within the dataset may exhibit different characteristics and behaviors. For instance, when analyzing COVID-19 data across regions, urban and rural areas may have distinct patterns. Handling this heterogeneity while preserving the interdependencies between entities is a complex task.

- **Data Imbalance:** In certain applications, some entities or regions may have significantly more data points or observations than others. Data imbalance can introduce bias and affect the model's ability to generalize to less represented entities.

- **Computation and Resource Intensiveness:** Processing high cardinality multivariate multi-time series datasets often requires substantial computational resources. Advanced modeling techniques, such as deep learning, can be computationally intensive, and managing large-scale data efficiently is a challenge.

1.1.4.2 Importance of HCMVMT Dataset in the Context of Thesis

The significance of understanding and effectively modeling high cardinality multivariate multi-time series datasets becomes particularly pronounced when considered in the context of this thesis. The thesis, as previously detailed, is designed to advance the field's approach to predictive analytics in the face of such complex datasets and, specifically, to provide accurate predictions of epidemic trends at regional or provincial levels.

Localized epidemic trend forecasts are of paramount importance in situations like the COVID-19 pandemic, where a granular understanding of the data is essential for effective decision-making. A broad, national-level analysis often falls short of providing the precise insights needed for region-specific strategies and policies. It is at this juncture that the complexities of high cardinality multivariate multi-time series datasets are thrust into the spotlight.

The data analyzed in this thesis, which encompasses multiple regions or provinces, each with its unique set of variables, is a quintessential example of high cardinality multivariate multi-time series data. It exhibits the intertwined nature of variables, dynamic temporal dependencies, and the impact of real-world uncertainties. Effectively addressing these complexities, as the proposed framework aims to do, becomes pivotal in the journey to provide reliable and targeted predictions, thus informing decisions that have far-reaching implications.

1.1.5 Limitations of State-of-the-Art Methods for HCMVMT Dataset:

The state-of-the-art methods discussed in the previous section offer powerful tools for time series prediction. However, when applied to high cardinality multivariate multi-time series datasets, they face a series of limitations that hinder their effectiveness. These limitations must be considered when choosing an appropriate approach for addressing the complexities of such datasets.

1.1.5.1 Deep Neural Networks (DNNs)

Deep Neural Networks (DNNs)[30], while highly versatile, exhibit limitations when applied to high cardinality multivariate multi-time series datasets:

- **Incompatible with Raw Data:** DNNs may excel at capturing the total target value over time but might struggle to efficiently capture changes in the target value. This limitation can impact the modeling of dynamic trends and dependencies within the data.
- **Challenges in Capturing Change:** DNNs may excel at capturing the total target value over time but might struggle to efficiently capture changes in the target value. This limitation can impact the modeling of dynamic trends and dependencies within the data.
- **Dependency Handling:** DNNs, despite their depth and complexity, may not efficiently capture dependencies across previous time series steps. This is a critical limitation when dealing with time series data that exhibits intricate temporal relationships.

1.1.5.2 ARIMA Based Models

ARIMA (AutoRegressive Integrated Moving Average) [19] models, while effective for simpler time series, have specific limitations when handling high cardinality multivariate multi-time series datasets:

- **Univariate Limitation:** ARIMA models are inherently designed for univariate time series and may not directly handle multiple variables. This limitation complicates their application to high cardinality datasets characterized by multiple interacting variables.

- **Complex Dependency Handling:** These models assume that the current value can be predicted based on a linear combination of its previous values. In scenarios where complex dependencies exist between multiple time series, this assumption may not hold, impacting predictive accuracy.

- **Multivariate Extensions:** Multivariate versions of ARIMA models exist, but their applicability and performance can be constrained by the complexity and non-linearity of interactions between variables.

1.1.5.3 Vector Autoregression (VAR)

Vector Autoregression (VAR)[24], a multivariate extension of ARIMA, faces its own set of limitations when applied to high cardinality multivariate multi-time series datasets:

- **Linear Assumption:** Like ARIMA, VAR models also assume that the current value can be predicted based on a linear combination of its previous values. In cases where non-linear relationships exist, the models may not effectively capture these dynamics.

- **Performance with Many Variables:** The performance of VAR models can degrade when handling a high number of variables, as the complexity of interactions increases. This limitation poses challenges when dealing with high cardinality datasets with numerous variables.

- **Nonlinear Relationships:** VAR models are primarily designed for linear relationships between variables. In high cardinality datasets, nonlinear relationships may dominate, impacting the model's predictive accuracy and reliability.

1.1.5.4 Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs)[23], with their temporal modeling capabilities, offer advantages for time series data. However, they encounter specific limitations when applied to high cardinality multivariate multi-time series datasets:

- **Difficulty in Capturing Long-Term Dependencies:** RNNs may struggle with capturing long-range dependencies in time series data. While they have memory mechanisms, the vanishing gradient problem can hinder their ability to efficiently capture long-term patterns.
- **Computationally Intensive:** Training deep RNNs, especially on high-dimensional and high cardinality datasets, can be computationally intensive. Managing the resources required for these models is a practical challenge.
- **Data Pre-processing and Feature Engineering:** RNNs often require extensive data pre-processing and feature engineering, including scaling and normalizing the data, which can add complexity to the modeling process.
- **Handling Irregular Time Intervals:** When dealing with datasets containing irregular time intervals, RNNs may face difficulties. Ensuring that the model effectively handles gaps and irregularities in the data can be challenging.
- **Scalability:** RNNs might face challenges in scaling effectively to generalize well for different distribution series, especially in high cardinality datasets where numerous entities exhibit varying behavior.

1.1.5.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN)[22], a proximity-based algorithm, offers simplicity and interpretability. However, it exhibits several limitations when applied to high cardinality multivariate multi-time series datasets:

- **Computationally Expensive:** KNN can be computationally expensive, particularly when dealing with large datasets and higher values of k (number of

neighbors to consider). This computational intensity may hinder practical applications.

- **Sensitive to k:** The choice of k is a critical hyperparameter in KNN. Selecting the right value of k is essential for the model's performance, and it may require experimentation, which can be time-consuming.

- **Capturing Long-Term Dependencies:** KNN struggles with capturing long-term dependencies in time series data, particularly when patterns extend over multiple time steps. It excels at capturing local patterns but may miss global trends.

- **Incompatible with Raw Data:** KNN typically cannot be applied directly to raw, unprocessed time series data. Data preprocessing and transformation are often necessary, making the workflow more intricate.

1.2 Problem Definition

Let $X \in \mathbb{R}$ represent the feature space, where $R = \{X_{i,j}\}$ for $i \in [1, n]$, $j \in [1, m]$.

Let Z be the set of target variables, where $Z = y_i$.

Our objective is to establish a prediction framework with the purpose of training a machine learning function $F_m : X \to y$, such that $F_m(X_{i,j}) = y_i$.

The feature space X is characterized by $i \times j$ feature vectors $X_{i,j}$, where $i \in [1, n]$ represents the data samples in the dataset, and $j \in [1, m]$ represents the independent variables or features.

The target variables y_i correspond to the data samples in the dataset, with n denoting the total number of samples and m representing the number of independent variables in the dataset.

The function F_m strives to learn patterns within the information contained in X in order to provide accurate predictions for y based on this information.

For example, y_i could represent the number of cases, deaths, or hospitalizations.

X_i might correspond to a single timestamp or data point for which the target variable needs to be predicted.

X_j could represent the percentage of vaccination coverage for a specific dose (e.g., dose 1).

1.3 Problem Motivation

The endeavor to predict and understand COVID-19 epidemic trends, encompassing critical aspects such as the number of cases, deaths, and hospitalizations, has been a profound motivation guiding the research journey embarked upon in this thesis. This motivation arises from a recognition of the pressing need to address complex challenges and make informed decisions in the face of a global health crisis. The core motivations behind the choice to focus on provincial-level epidemic prediction are multi-fold, and they align closely with the essence of this thesis.

1.3.1 Beyond Limitations Of Previous Works

The motivation behind this research extends far beyond simply addressing the limitations of previous works in dealing with the prediction of epidemic trends. While recognizing the shortcomings of existing approaches, our motivation encompasses a broader scope, driven by a multitude of factors that underpin the relevance and significance of this work.

This motivation revolves around the exploration of three major questions:

1. How does the COVID-19 problem and its corresponding dataset relate to the concept of High Cardinality Multivariate Multi-time Series (HCMVMT) datasets?

2. Why should COVID-19 be dealt with in a manner that leads to the creation of HCMVMT datasets?

3. What are the advantages of the results obtained when COVID-19 is approached as an HCMVMT dataset?

The research conducted in this thesis aims to provide comprehensive answers to these questions, shedding light on the deeper connections between COVID-19 and

HCMVMT datasets, the rationale for approaching COVID-19 in this unique manner, and the tangible benefits that arise from this novel perspective. By doing so, it transcends the limitations of previous works and underscores the holistic approach taken in addressing the complexities of epidemic trend prediction.

1.3.2 High Cardinality Multivariate Multi-time Series Datasets and COVID-19

Predicting COVID-19 epidemic trends at the provincial level leads to the generation of such high cardinality multivariate multi-time series datasets. Each province's data represents a unique, evolving time series characterized by multiple variables. The complexity arises from the interplay of these variables over time, making the dataset intricate and challenging to analyze.

This complexity and the scale of data align directly with the motivation of this research. The understanding is that addressing the challenges posed by high cardinality datasets, where each region's dynamics are influenced by numerous interrelated factors, is a pivotal step in advancing the field of predictive analytics. It is within this intricate web of data that critical insights are waiting to be discovered.

The thesis's objective is to introduce a framework tailored to deal with the complexities of these high cardinality multivariate multi-time series datasets. This framework aims to provide reliable predictions by modeling the dependencies, interactions, and temporal dynamics inherent in such data. By focusing on this unique dataset structure, the research aims to bridge the gap between the complexity of real-world epidemic trends and the capabilities of data analysis and prediction, highlighting the intricate nature of the data and its significance in informing effective decision-making at the provincial, national, and international levels.

1.3.3 The Significance of Provincial-Level Predictions

The importance of predicting COVID-19 at the provincial level extends to numerous critical aspects of managing a pandemic. This approach recognizes the dynamic and

diverse nature of the COVID-19 crisis and seeks to address these challenges with a localized, nuanced perspective. Several key reasons underscore the significance of provincial-level predictions:

1.3.3.1 Localized Decision Making

Different provinces or regions may experience varying infection rates, healthcare capacities, and vaccination coverage. Models that offer predictions at the provincial level allow policymakers and public health officials to make more informed decisions tailored to the specific needs of each region. This approach can help optimize resource allocation and response strategies based on the unique circumstances in each province.

1.3.3.2 Targeted Interventions

Understanding the potential trajectory of the virus at the province level enables authorities to implement targeted interventions and containment measures in areas that are likely to be most affected. This proactive approach can help prevent widespread outbreaks and mitigate the impact on public health and the economy. By identifying hotspots early, authorities can deploy resources more efficiently.

1.3.3.3 Resource Planning

Healthcare systems may vary significantly from province to province, both in terms of capacity and readiness. Predictive models at the province level can aid in estimating the demand for medical resources, such as hospital beds, ventilators, and medical personnel. This, in turn, allows for better preparedness and resource allocation. Knowing which provinces are at higher risk of surges in cases enables healthcare systems to plan and allocate resources strategically.

1.3.3.4 Risk Assessment

Provincial-level predictions provide valuable insights into which areas are at higher risk of experiencing surges in cases. This information can be used to prioritize surveil-

lance, testing, contact tracing, and vaccination efforts in regions that are most vulnerable. Risk assessment at the province level enables a more targeted and effective public health response.

1.3.3.5 Monitoring and Evaluation

Comparing model predictions with actual outcomes at the province level is crucial for continuous evaluation. Health authorities can assess the accuracy of their models and improve their forecasting capabilities over time. This iterative process helps refine strategies and responses as the pandemic evolves, ensuring that interventions remain effective.

1.3.3.6 Research and Collaboration

COVID-19 models at the provincial level can foster research collaborations and information sharing between provinces and countries. Scientists and health experts can learn from each other's experiences and adapt successful strategies to their own regions. This collaborative approach accelerates the development of effective responses to the pandemic.

In essence, predicting COVID-19 at the provincial level is not merely an analytical exercise but a strategic imperative. It acknowledges the diverse nature of the pandemic and emphasizes the importance of localized decision making, targeted interventions, resource planning, risk assessment, monitoring and evaluation, research collaboration, and information sharing. By understanding the unique dynamics within each province, we can tailor responses and interventions to effectively combat the virus, ultimately reducing its impact on public health and the economy. This approach offers a comprehensive and localized solution to managing the complexities of the COVID-19 pandemic.

1.3.4 Abstraction to National and International Levels

Dealing with the COVID-19 pandemic at the provincial level and generating predictions for each province brings with it a unique advantage that extends well beyond local response. The results obtained from provincial-level predictions can be efficiently abstracted to the national and international levels, facilitating a seamless and accurate transfer of insights. This abstraction process is underpinned by several key factors:

1.3.4.1 Data Granularity

The provincial-level approach provides a high degree of data granularity. Each province is considered individually, accounting for the specific characteristics and conditions within that region. This granularity allows for a detailed understanding of how the virus behaves within different settings and under varying circumstances. As a result, the insights gained at the provincial level are highly granular and specific.

1.3.4.2 Aggregation and Summation

One of the advantages of working with granular data is the ability to aggregate and sum the results. Aggregation involves combining the data and insights from individual provinces to create a broader view. This process is relatively straightforward, as it involves summing up the predictions, metrics, and findings from each province to create a national-level perspective.

1.3.4.3 Comparisons and Contrasts

The granular data from provincial-level predictions also enables meaningful comparisons and contrasts. Health authorities, policymakers, and researchers can readily compare the experiences and responses of different provinces. These comparisons are valuable for identifying best practices and strategies that have been particularly effective in one region and may be applicable elsewhere.

1.3.4.4 Hierarchical Modeling

Hierarchical modeling is a powerful tool that leverages provincial-level data to build models that can be applied at national or international levels. By understanding the relationships between provinces and the factors that influence the virus's spread, hierarchical models can provide insights at various scales. Provincial-level data serves as a foundational element in constructing these models.

1.3.4.5 Fine-Tuning and Generalization

The provincial-level data provides a robust foundation for fine-tuning models and generalizing findings. It allows for the refinement of predictive models and strategies at the provincial level. Once these models are fine-tuned and proven effective, they can be scaled up and applied to national and international scenarios with greater confidence.

1.3.4.6 Policy and Strategy Sharing

Provincial-level predictions are a valuable resource for sharing policies and strategies that have yielded positive outcomes. By understanding what has worked in one province, policymakers in other regions can adopt similar measures. This sharing of successful strategies can be highly beneficial on a larger scale.

1.3.4.7 Consistency and Standardization

Working at the provincial level promotes consistency and standardization in data collection, reporting, and response strategies. This uniformity facilitates the creation of a cohesive and standardized approach to the pandemic. When the same procedures are followed across provinces, the data generated is more easily integrated and compared.

In essence, dealing with COVID-19 at the provincial level not only offers a localized and targeted response but also simplifies the process of abstracting results to national and international levels. The granularity of data, ease of aggregation, poten-

tial for hierarchical modeling, and opportunities for policy and strategy sharing make the transition from provincial insights to broader applications smoother and more effective. This hierarchical approach, driven by granular data and localized decision-making, ensures that the knowledge and insights gained are readily applicable on a larger scale, ultimately contributing to more effective pandemic management at the national and international levels.

1.4 Thesis Statement

This research introduces a novel and innovative hybrid framework designed to address the intricate challenges posed by High Cardinality Multivariate Multi-time Series (HCMVMT) datasets, with the overarching objective of achieving improved predictive performance. The proposed framework is meticulously crafted to excel on key performance metrics, including R^2 (Coefficient of Determination), RMSE (Root Mean Square Error), Mean Absolute Error (MAE), and Max Absolute Error (Max AE).

To fulfill this objective, a practical and meticulously constructed COVID-19 dataset, enriched with HCMVMT characteristics, is developed. This dataset serves as the foundation for predicting epidemic trends at the provincial level. The incorporation of HCMVMT features allows for a comprehensive understanding of the multifaceted factors influencing epidemic dynamics.

Central to this research is a rigorous comparative analysis. The hybrid framework is rigorously evaluated against individual state-of-the-art methods to ascertain its effectiveness in enhancing predictive performance across the specified metrics. By directly contrasting the novel hybrid framework with established methodologies, this study provides critical insights into its strengths and its capacity to outperform existing approaches.

Through the development of this hybrid framework and the creation of a practical COVID-19 dataset, this research advances the field of predictive analytics, particularly in the context of high cardinality multivariate multi-time series datasets. It not only showcases the framework's potential to tackle the complexities of such data

but also underscores its practical significance in informing critical decisions for managing epidemic trends at the provincial level. The findings of this study hold the promise of offering more accurate and targeted predictions, thus contributing to effective decision-making in healthcare, public policy, and beyond.

1.5 Thesis Objectives and Contributions

This section outlines the objectives of this thesis and the contributions it makes to the field of predictive analytics. The primary objectives of this research encompass the following key areas:

1.5.1 Comprehensive Canadian Dataset

One of the core objectives of this research is the development of a Comprehensive Canadian Dataset tailored for predicting epidemic trends. This dataset goes beyond traditional data sources, incorporating critical features related to COVID-19 and its impacts on the Canadian provinces. It serves as a foundational resource for understanding and predicting the multifaceted dynamics of the pandemic at the provincial level. The dataset aims to capture the complexities of high cardinality multivariate multi-time series datasets, enabling a nuanced analysis of epidemic trends.

1.5.2 Framework to Deal with Multivariate Multi-timeseries Datasets

A fundamental contribution of this research is the introduction of a novel and innovative Framework specifically designed to address the intricacies of multivariate multi-time series datasets. This framework seamlessly integrates advanced data preprocessing, feature selection, feature engineering, feature encoding, and model architecture. It is meticulously crafted to capture the dependencies, interactions, and temporal dynamics inherent in complex datasets. By providing a comprehensive solution to the challenges of such datasets, this framework advances the field of predictive

analytics and opens avenues for robust predictions across various domains.

1.5.3 Comprehensive Single AI Model for Epidemic Trends

The development of a Comprehensive Single AI Model represents a significant milestone in this research. The objective is to create a unified model that performs exceptionally well in predicting various epidemic trends, such as the number of cases, deaths, and hospitalizations. This single model is designed to work effectively for all epidemic trends, streamlining the prediction process and ensuring consistent accuracy. It simplifies the predictive analytics workflow and provides a versatile tool for decision-makers and researchers.

1.5.4 Critical Feature Identification

Identification of Critical Features is another vital objective of this research. By discerning the most influential variables and factors in the context of epidemic trends, the research aims to provide decision-makers with actionable insights. The identification of these features contributes to the development of more precise and effective predictive models. It also enhances the understanding of the key determinants shaping epidemic trends.

In summary, this thesis sets forth a multifaceted agenda aimed at advancing the field of predictive analytics, particularly in the context of high cardinality multivariate multi-time series datasets. The objectives outlined here, including the creation of a comprehensive Canadian dataset, the development of a novel framework, the establishment of a single AI model for epidemic trends, and the identification of critical features, collectively contribute to the enhancement of predictive analytics and its practical significance in informing critical decisions for managing epidemic trends at the provincial level. The outcomes of this research hold the potential to transform the way we approach predictive analytics in the face of complex, real-world data challenges.

1.6 Thesis Organization

The structure of this research/thesis work is outlined as follows:

In Chapter 2, we embark on an exhaustive exploration of previous research endeavors within the realms of predicting COVID-19 epidemic trends, delving into both regression and time series methodologies, while also examining their inherent constraints. Additionally, we engage in a critical discussion regarding the limitations of contemporary methodologies and models, particularly focusing on their efficacy when confronted with the challenges posed by High Cardinality Multivariate Multi-Time Series (HCMVMT) datasets.

In Chapter 3, we will explore critical concepts encompassing feature engineering, selection, dimension reduction techniques, deep learning and machine learning models, as well as the evaluation metrics employed. These elements collectively serve as the bedrock for our framework, which extends its utility beyond the prediction of COVID-19 epidemic trends to offer a comprehensive solution for HCMVMT datasets across diverse domains.

In Chapter 4, we embark on an exploratory journey into the heart of our research, where we delve into the intricacies of data, data handling, and the foundations upon which our analysis is built. We introduce a novel dataset, meticulously collected over an extensive timeframe, serving as the bedrock for our investigations. The chapter unravels the nuances of data collection and description, providing a comprehensive overview of the wealth of information encapsulated within. Furthermore, we address the challenges posed by missing values and null entries, employing innovative techniques to ensure the completeness of our dataset. We also unveil our approach to data versioning, leading us to the discovery of the best-performing dataset. These critical components are the cornerstone of our analytical journey, paving the way for profound insights and impactful conclusions.

In Chapter 5, we unveil a unified framework that seamlessly combines High Cardinality Multivariate Multi Time Series (HCMVMT) datasets and COVID-19 epidemic trend prediction. Recognizing the inherent similarities in these domains, the frame-

work is divided into two dimensions: one focusing on regression-based forecasting and the other on time series modeling. This comprehensive approach addresses the distinct challenges of both realms while fostering interdisciplinary insights, ultimately elevating the accuracy of our predictions.

In Chapter 6, we delve into the meticulous details of the Experimentation Setup for both the Regression Framework and the Deep Learning Time Series Framework. We outline the specific configurations and parameters employed in each framework, providing transparency into the choices made during the experimentation process. Following this, we elucidate the comprehensive Evaluation Metrics utilized to assess the performance of these frameworks, covering critical aspects such as R-squared, Mean Absolute Error, Max Absolute Error, and Root Mean Squared Error. The subsequent section unveils the Results, encompassing the identification of critical features in the Regression Model and the detailed outcomes of the Time Series Models, including specific configurations for Prophet, LSTMs[20], GRU[14], and DeepAREstimator[25]. The Discussions section critically examines the statistical stability and reliability of the obtained results, elucidates the assumptions inherent in both frameworks, delineates the limitations faced during experimentation, and culminates with a reflection on the contributions made by this research to the domain of epidemic trend prediction.

In Chapter 7, we encapsulate the journey of the Regression Framework and the Deep Learning Time Series Framework, offering a comprehensive synthesis of their performances, trade-offs, and advantages. Delving into the Regression Framework, we examine its unparalleled outperformance in prediction metrics, showcasing its superiority over traditional models. However, a notable trade-off surfaces in the form of extended training times, posing practical challenges. The framework's scalability and maintainability shine as pivotal achievements, streamlining model management and reducing complexity. Transitioning to the Deep Learning Time Series Framework, we dissect the performance metrics, emphasizing the noteworthy outperformance of Time Series Models over statistical counterparts. While a trade-off emerges in terms of training time, the introduction of the DeepAREstimator model offers a promising compromise, maintaining competitive performance with reduced training times. The

chapter concludes by navigating the terrain of selecting the optimal solution based on specific priorities, emphasizing practical considerations such as time, performance, and sustainability. Looking ahead, the chapter segues into Future Works, charting a course for expanding datasets to a global scale, experimenting with alternative techniques, exploring structural transformations like knowledge graphs, and embracing advanced machine learning models. This exploration underscores the continuous evolution and potential enhancements awaiting these frameworks in the realm of epidemic trend prediction.

CHAPTER 2

Related Works

In the realm of predicting COVID-19 epidemic trends, a considerable body of research has emerged, reflecting the collective efforts to decipher the complex dynamics of this global health crisis. These previous works have sought to harness the power of data-driven methodologies to anticipate the evolution of the pandemic and inform effective public health strategies. While these endeavors have undeniably contributed valuable insights, it is essential to critically evaluate their methodologies, findings, and inherent limitations. This section embarks on a comprehensive exploration of prior research, delving into the methodologies employed, the gaps they have left unaddressed, and the constraints that have hindered their efficacy. By recognizing the shortcomings of these earlier attempts, we can pave the way for a more robust and accurate approach to predicting COVID-19 epidemic trends.

In the context of predicting the complex and dynamic patterns of the COVID-19 epidemic, researchers and data scientists have predominantly gravitated toward two distinct yet complementary methodological avenues. These approaches, namely regression models and time series models, have emerged as the principal pillars of analysis in the quest to forecast and comprehend epidemic trends. On one hand, regression models offer the power of statistical inference, enabling the identification of significant factors influencing the spread and impact of the virus. By establishing relationships between independent variables and epidemic outcomes, these models facilitate the quantification of the effects of interventions, socio-economic determinants, and other crucial drivers. On the other hand, time series models specialize in capturing the intricate temporal dependencies that characterize the evolution of the

pandemic. Leveraging these models, researchers can account for seasonality, trends, and autocorrelation in the data, thus providing a detailed understanding of how the epidemic evolves over time. While each of these methodologies has its unique strengths, they also exhibit inherent limitations, such as the potential oversimplification of the problem in regression models or the complexity of fine-grained temporal modeling in time series approaches. This section endeavors to scrutinize the successes and constraints of these two major approaches, shedding light on the nuanced interplay between them in the pursuit of accurate and comprehensive predictions of COVID-19 epidemic trends.

2.1 Review of Models for Predicting COVID-19 as Classification Problem

Zhang et al. [12] applied a Support Vector Machine (SVM) model for COVID-19 cases detection and classification. The clinical information and blood/urine test data were used in their work to validate SVM's performance. Simulation results demonstrated the effectiveness of the SVM model by achieving an accuracy of 81.48%, sensitivity of 83.33%, and specificity of 100%.

Sun et al. [5] used SVM model for predicting the COVID-19 patients with severe/critical symptoms. 220 clinical/laboratory observations records and 336 cases of patients infected COVID-19 divided into training and testing datasets were used to validate the performance of the SVM model. Simulation results showed that the SVM model achieves an Area Under Curve (AUC) of 0.9996 and 0.9757 in the training and testing dataset, respectively.

Nour et al.[6] applied machine learning approaches such as SVM, Decision tree (DT), and KNN for automatic detection of positive COVID-19 cases. The performance of the proposed approaches was validated on a public COVID-19 radiology database divided into training and test sets with 70% and 30% rates, respectively.

Tabrizchi et al. [35] used SVM with Naive Bayes (NB), Gradient boosting decision

tree (GBDT), AdaBoost, CNN, and Multilayer perceptron (MLP) for rapid diagnosis of COVID-19. A dataset of 980 CT scan images (430 with COVID-19 and 550 normal) was used in the simulation, and results showed that SVM outperforms other machine-learning approaches by achieving an average accuracy, precision, sensitivity, and F1-score of 99.20%, 98.19%, 100%, and 99.0%, respectively.

2.2 Review of Regression Models for Predicting COVID-19 Epidemic Trends

Predicting the trends of the COVID-19 epidemic is a challenging task with far-reaching implications. Various research studies have attempted to tackle this problem through regression models. In this section, we review the existing literature and highlight the key limitations and gaps in previous works.

2.2.1 Utilization of Traditional Machine Learning Methods

Yue et al. [4] utilized a linear regression model for predicting COVID-19 infected patients. CT images of 52 patients from five hospitals (Ankang, Lishui, Zhenjiang, Lanzhou, and Linxia) were used to assess the performance of the regression model. Simulation results demonstrated that the linear regression model outperforms the Random Forest algorithm.

Salama et al. [1] employed the linear regression model with SVM and ANN for predicting COVID-19 infected patients. The proposed models were assessed using an Epidemiological dataset collected from real-time health reports. Simulation results indicated that SVM had the lowest mean absolute error (0.21), while the regression model had the lowest root mean squared error (0.46).

Yadav et al. [7] used three machine learning approaches (Linear Regression, Polynomial Regression, and SVR) for COVID-19 epidemic prediction and analysis. The dataset included the total number of COVID-19 positive cases from various countries, and results showed the superiority of SVR compared to Linear Regression and

Polynomial Regression, with average accuracies of 99.47%, 65.01%, and 98.82%, respectively.

Khanday et al. [3] proposed Logistic Regression with six machine learning approaches (Adaboost, Stochastic Gradient Boosting, Decision Tree, SVM, Multinomial Naïve Bayes, and Random Forest) for COVID-19 detection and classification. Evaluation with 212 clinical reports divided into four classes (COVID, ARDS, SARS, and Both) showed that logistic regression provided excellent performance, with 94% precision, 96% sensitivity, an accuracy of 96.20%, and a 95% F1-score.

Saqib [2] developed a novel model (PBRR) by combining Bayesian Ridge Regression (BRR) with an n-degree Polynomial for forecasting COVID-19 outbreak progression. The PBRR model's performance was validated using public datasets collected from John Hopkins University available until May 11, 2020. Experimental results revealed the good performance of PBRR, with an average accuracy of 91%.

Most of the current regression-based studies have predominantly relied on traditional machine learning methods. These methods include k-Nearest Neighbors (KNN) and Random Forest, which are powerful techniques but may not fully capture the complexity of epidemic trends. While these approaches have provided valuable insights, they often lack the ability to uncover intricate relationships within the data.

2.2.2 Limited Exploration of Deep Learning Techniques

Chimmula et al. [8] employed an LSTM model to predict COVID-19 cases in Canada. The effectiveness of the LSTM model was verified using data from Johns Hopkins University and the Canadian Health Authority, encompassing numerous confirmed cases. The outcomes revealed that the LSTM model demonstrated superior performance compared to alternative forecasting models.

A solitary study, conducted by [26], took a Multi-Factor Deep Learning approach to predict COVID-19 epidemic trends. However, this study used a limited set of features and achieved a relatively low R^2 value (0.65) and less-than-optimal Root Mean Square Error (RMSE) and Mean Squared Error (MSE) metrics. This highlights the need for more comprehensive and accurate modeling techniques to better understand

and predict epidemic dynamics.

2.2.3 Issues with Normalization

One common issue observed in previous work [26] is the incorrect normalization of data. This misstep can lead to the inadvertent introduction of correlations between features by a factor of $-\frac{1}{p-1}$, where p represents the number of features[28]. Such normalization errors can significantly impact the validity of regression models, emphasizing the importance of proper data preprocessing.

2.2.4 Country-Level Models

Many prior studies have developed regression models at the national or country level. While these models provide valuable insights into COVID-19 trends on a broad scale, they are often insufficient for addressing specific regional or provincial concerns. The effectiveness of preventive actions, which can be crucial for halting the spread of the disease, may vary significantly across different regions.

2.2.5 Lack of Identifiable Epidemic Trend Identifiers

Previous research endeavors have struggled to identify and confirm critical factors or variables that significantly influence epidemic trends. This limitation hinders our ability to develop accurate predictive models and implement targeted interventions effectively.

2.2.6 Generalizability Across Regions and Epidemic Trends

Perhaps one of the most significant shortcomings of previous studies is the lack of a single, universally applicable model or architecture. These models often fail to generalize well across various regions, encompassing provinces, states, and different countries, and for all types of epidemic trends. The absence of a versatile framework hinders the development of comprehensive epidemic forecasting models.

In light of these limitations and gaps in existing literature, this thesis aims to address these challenges and contribute to the development of more robust and accurate regression models for predicting COVID-19 epidemic trends.

2.3 Review of Time Series Models for Predicting COVID-19 Epidemic Trends

Previous research endeavors have delved into the formidable challenge of predicting COVID-19 epidemic trends by primarily adopting time series models. These studies, while valuable in their contributions, exhibit several common limitations that have hindered the development of comprehensive and versatile predictive models.

2.3.1 Absence of Multi-Factor Time Series Approaches

Remarkably, none of the previous studies have explored the application of a multi-factor time series approach or a multi-factor multi-time series approach to address the intricate dynamics of COVID-19 epidemic trends. The absence of such multi-dimensional modeling is a notable limitation, as it overlooks the potential interactions and dependencies between multiple epidemic drivers, making it challenging to uncover the complete causal web governing the pandemic's progression. A comprehensive model that integrates a multitude of relevant factors could offer a more nuanced understanding and accurate predictions of COVID-19 trends.

2.3.2 Overreliance on ARIMA Models

Chakraborty and Ghosh [9] devised a hybrid approach (ARIMA–WBF) that combines the ARIMA model with Wavelet-based forecasting (WBF) to predict the daily confirmed COVID-19 cases. The efficacy of ARIMA-WBF was verified using datasets comprising 346 cases from five countries (Canada, France, India, South Korea, and the UK). Simulation results underscored the effectiveness and robustness of the ARIMA-WBF method in forecasting COVID-19 cases.

Singh et al. [10] proposed Least Square-SVM (LS-SVM) and Autoregressive Integrated Moving Average (ARIMA) for the prediction of COVID-19 cases. A dataset of COVID-19 confirmed cases collected from five the most affected countriesFootnote1 was used to validate the proposed models. It was demonstrated that the LS-SVM model outperforms the ARIMA model by obtaining an accuracy of 80

Ribeiro et al. [11] conducted a comparable study, employing six machine learning methodologies, including stacking-ensemble learning (SEL), support vector regression (SVR), cubist regression (CUBIST), auto-regressive integrated moving average (ARIMA), ridge regression (RIDGE), and random forest (RF), to predict outcomes in COVID-19 datasets.

A pervasive trend among prior studies is the extensive use of Autoregressive Integrated Moving Average (ARIMA) based time series models. While ARIMA models are a foundational tool in time series analysis, they come with inherent limitations. These models assume linear relationships between variables, which may not adequately capture the complex, often nonlinear, interactions within epidemic data. Furthermore, ARIMA models struggle to capture long-term dependencies and intricate patterns that extend beyond the chosen lag order, potentially leading to suboptimal forecasting accuracy. Moreover, ARIMA models primarily focus on the time series itself and do not explicitly incorporate external factors or predictors that could be crucial in understanding and predicting COVID-19 dynamics.

2.3.3 Country-Level Predictions with Limited Generalizability

Most of the previous studies have focused on producing time series models for country-level COVID-19 predictions. While these models provide valuable insights into the overall trends, they often lack the versatility required for more localized decision-making and practical applications. The effectiveness of public health interventions and containment strategies can vary significantly at regional or provincial levels, and these country-level models may not sufficiently address such variations. As a re-

sult, there remains a considerable gap in producing models that generalize well for provinces, states, and countries, accommodating diverse epidemic trends and regional disparities.

2.3.4 Lack of a Unified, Generalizable Model

Perhaps the most critical limitation observed in previous research is the failure to develop a single unified model or architecture that can generalize effectively across regions, provinces, and countries and for all types of epidemic trends. The absence of such a versatile model hinders the development of comprehensive predictive tools that can support decision-makers, healthcare professionals, and policymakers across the globe.

This section underscores the common limitations in existing research, highlighting the opportunities for more innovative and comprehensive time series modeling in the prediction of COVID-19 epidemic trends. The following sections of this thesis seek to address these limitations and contribute to the development of more accurate and adaptable forecasting models.

2.4 Review of Datasets Used for Predicting COVID-19 Epidemic Trends

The datasets employed in previous research on predicting COVID-19 epidemic trends have played a pivotal role in shaping the capabilities and limitations of the models developed. While these studies have contributed valuable insights, they exhibit several significant limitations with regard to dataset utilization.

2.4.1 Limited Use of the John Hopkins Dataset

A recurring trend in many of the previous studies is the predominant reliance on the John Hopkins Dataset [16] as the primary source of data for predicting COVID-19 epidemic trends. However, a key limitation of this approach is that it often involves

using this dataset without incorporating additional relevant features. Many of these studies have focused primarily on predicting the number of cases and deaths, thus limiting their scope and ability to capture the multifaceted dynamics of the pandemic. The overreliance on this dataset restricts the exploration of a more comprehensive set of factors that could enhance the accuracy of predictions.

2.4.2 Predictions at the National Level

Another limitation of relying solely on the John Hopkins Dataset is that it often leads to predictions at the national level. Such predictions, while valuable for an overall understanding of the pandemic's trajectory, may lack practical applicability at more localized levels. The effectiveness of public health interventions and containment strategies can vary significantly between provinces, states, and regions, and national-level models may not capture these variations adequately.

2.4.3 Sparse Consideration of Multi-Factor Features

Among the multitude of previous works, a solitary study [26] stands out for its attempt to consider multiple feature groups, including environmental factors, human factors, biological factors, and government actions. However, this endeavor is constrained by several limitations. Notably, the study incorporates only a limited set of concrete features (13) within these feature groups, falling short of fully capturing the complex interplay of variables that influence the spread of COVID-19. Furthermore, the data normalization technique applied is detrimental to the robustness and interpretability of the models, potentially undermining the quality of predictions.

2.4.4 Opportunities for More Comprehensive Feature Groups

Emerging state-of-the-art studies focusing on the spread of COVID-19 underscore the potential for utilizing a broader spectrum of feature groups. These studies have indicated that incorporating factors such as vaccination coverage, mobility patterns, age distribution, geographical and land characteristics, economic factors, medical char-

acteristics, health characteristics, disease prevalence, and population behaviors can lead to more accurate and informative models for predicting not only the number of cases and deaths but also hospitalizations. This comprehensive approach acknowledges the multi-faceted nature of the pandemic and the need to capture the complex interdependencies between diverse factors.

2.4.5 Absence of Holistic Feature Integration

Despite the recognition of the importance of diverse feature groups, no previous study has undertaken the ambitious task of considering all the concrete factors belonging to these essential groups for predicting COVID-19 epidemic trends. This gap highlights a critical limitation in the existing body of research and emphasizes the untapped potential for the development of holistic, all-encompassing predictive models.

In light of these limitations associated with dataset utilization, this thesis aims to address these shortcomings by comprehensively integrating diverse feature groups and exploring the multifaceted dynamics of the COVID-19 pandemic. By doing so, we strive to enhance the accuracy and depth of predictive models, providing a more holistic understanding of COVID-19 epidemic trends.

2.5 Challenges of Handling High Cardinality Multivariate Multi-Time Series Datasets

High cardinality multivariate multi-time series (HCMVMT) datasets, characterized by a multitude of variables and intricate temporal dependencies, present a unique set of challenges for predictive modeling. Existing modeling techniques, particularly ARIMA and VAR-based models, often fall short when applied to these complex datasets.

2.5.1 Limitations of ARIMA Models

2.5.1.1 Linearity Assumption

ARIMA models, which have been widely employed in time series analysis, are based on the assumption that the relationships between variables are linear. This assumption may not hold true for HCMVMT datasets, which can exhibit complex nonlinear relationships. The linearity assumption can lead to potential model misspecification, resulting in inaccurate predictions.

2.5.1.2 Inability to Incorporate External Factors

ARIMA models primarily focus on the time series itself and do not explicitly incorporate external factors or predictors. In the context of HCMVMT datasets, where multiple influential variables may exist, ARIMA models may not fully capture the effects of these factors on the time series. This can result in limited predictive accuracy, as important information remains unutilized.

2.5.1.3 Limited Lag Dependence

ARIMA models capture autocorrelation in the data by considering the lagged values of the series. However, they may not effectively capture long-term dependencies or complex patterns that extend beyond the chosen lag order. In HCMVMT datasets, where temporal dependencies can span a wide range of lags, this limitation can hinder the modeling of the underlying dynamics.

2.5.1.4 Long-Term Forecasting Limitations

ARIMA models are generally better suited for short- to medium-term forecasting. When it comes to long-term forecasting, they face challenges due to inherent uncertainty and potential structural changes in the data. This can limit the accuracy and reliability of ARIMA model predictions when applied to HCMVMT datasets.

2.5.2 Limitations of VAR Models

2.5.2.1 Nonlinear Relationships

Vector Autoregressive (VAR) models capture linear dependencies between variables but may struggle to capture complex nonlinear relationships present in HCMVMT datasets. The oversimplification of relationships between variables can result in a less accurate representation of the data.

2.5.2.2 Endogeneity Assumption

VAR models assume that the variables in the system are endogenous, meaning they depend on each other. In cases where external factors or exogenous variables play a significant role in HCMVMT datasets, the endogeneity assumption may not hold, leading to limited modeling flexibility and accuracy.

2.5.2.3 Curse of Dimensionality

As the number of variables in HCMVMT datasets increases, the number of parameters in a VAR model also increases. This phenomenon, known as the "curse of dimensionality," can make parameter estimation more challenging and render the model more susceptible to overfitting, especially when the dataset contains a limited number of observations.

2.5.2.4 Limitations for Non-Stationary Time Series

VAR models assume stationarity in the underlying data. However, HCMVMT datasets often consist of non-stationary time series, making VAR modeling directly inapplicable. This limitation can hinder the ability to capture the evolving dynamics of the data accurately.

In summary, the absence of a comprehensive framework for handling HCMVMT datasets and the limitations associated with ARIMA and VAR-based models have motivated the need for more advanced and adaptable

modeling techniques. This thesis seeks to address these challenges by developing novel approaches that can effectively handle the intricacies of HCMVMT datasets and provide accurate predictions of COVID-19 epidemic trends.

CHAPTER 3

Feature Selection, Extraction, DNN as Encoders and Important models

In this chapter, we embark on a journey into the fundamental underpinnings of our research, where we lay the essential groundwork for the comprehensive framework that follows. We delve deep into the realm of feature engineering, feature selection, dimension reduction techniques, and the eclectic array of machine learning models that constitute the heart of our analysis. These core concepts are the building blocks upon which our framework is constructed, providing a solid foundation for our predictive endeavors.

As we navigate through this chapter, we begin by elucidating the critical importance of feature engineering, where raw data is transformed into meaningful, informative features. We explore the nuances of feature selection, highlighting the art of choosing the most pertinent attributes to enhance model performance. Dimension reduction techniques, such as Principal Component Analysis (PCA) come into play as we seek to distill complex data into more manageable dimensions.

The chapter also serves as a canvas for the introduction of diverse deep learning models as part of our framework, offering a fresh perspective on problem-solving within the high cardinality, multivariate, multi-timeseries (HCMVMT) dataset domain. In place of classical models, we delve into the intricacies of deep learning, leveraging neural networks to tackle the multifaceted challenges presented by these

complex datasets. From recurrent neural networks (RNNs) for sequential data and beyond, we present a diverse array of deep learning architectures that empower us to address the intricate dynamics of high-dimensional, high-cardinality datasets.

Moreover, we shine a spotlight on the evaluation metrics that will be our guiding compass throughout the research journey, ensuring that our models are rigorously assessed against the highest standards.

As we traverse through this chapter, we invite you to join us on this journey of exploration and discovery. The foundations we lay here will serve as the solid bedrock upon which our methodology and framework are constructed. These are the building blocks that pave the way for a profound understanding of the intricate dynamics at play in our quest to predict and respond to the ever-evolving landscape of the COVID-19 pandemic.

As we embark on this educational expedition, it's crucial to recognize that the framework we construct extends beyond the realm of COVID-19 epidemic trend prediction. Its versatility and applicability transcend domains, offering a blueprint for the analysis of high cardinality, multivariate, multi-timeseries datasets across various fields. This framework is not only poised to enhance our understanding and prediction of pandemic trends but also holds the potential to illuminate the intricate dynamics of complex data in diverse disciplines, from healthcare to finance, and beyond. It is a powerful tool, poised to redefine the way we approach data-driven decision-making and predictive analytics, and it's our pleasure to take you on this enlightening journey.

3.1 Feature Selection - Mutual Information Regression (MIR)

Feature selection is a crucial step in data analysis and modeling, as it holds the potential to uncover the most informative attributes while discarding noise and redundancy. One prominent approach that has gained significant attention in recent years is Mutual Information Regression (MIR)[34]. In this writeup, we delve into the

workings of MIR[34], its advantages, and its remarkable effectiveness in identifying the most relevant features for predictive modeling.

3.1.1 Mutual Information Regression (MIR): A Primer

Mutual Information (MI) is a concept borrowed from information theory, which quantifies the dependence between two random variables. In the context of feature selection, MIR assesses the relationship between each feature and the target variable. It measures how much information about the target variable can be extracted from each feature. When MIR is applied to regression tasks, it helps us understand how well a feature can predict the target variable.

3.1.2 The Working Mechanism of MIR

MIR operates on the principle of information gain. It calculates the reduction in uncertainty about the target variable when the value of a particular feature is known. The higher the information gain, the more valuable the feature is in predicting the target.

To calculate the Mutual Information between a feature and the target variable, MIR evaluates the joint distribution of the two variables. In a regression context, this means measuring the dependency between the feature and the continuous values of the target variable. MIR computes the reduction in uncertainty of the target variable after considering the feature. This reduction in uncertainty, measured in bits, is the Mutual Information.

3.1.3 Advantages of Mutual Information Regression

MIR offers several advantages that make it a powerful tool in feature selection:

- **Non-linearity Tolerance:** One of the significant advantages of MIR is its ability to capture non-linear relationships between features and the target variable. Traditional linear methods may fail to recognize complex associations, but MIR can reveal them effectively.

- **No Assumption of Linearity:** MIR doesn't make the assumption of linearity, which is common in some feature selection techniques. It remains effective even when the relationship between features and the target is non-linear.

- **Robust to Irrelevant Features:** MIR tends to assign low Mutual Information values to irrelevant or noisy features. This robustness ensures that only the most informative features are selected, leading to more precise models.

- **Variable Selection:** MIR not only quantifies the importance of features but also performs variable selection. It identifies and ranks features according to their predictive power.

- **Feature Ranking:** MIR provides a ranking of features based on their Mutual Information with the target variable. This ranking is invaluable for understanding the relative importance of each feature.

3.1.4 Effectiveness of Mutual Information Regression

The effectiveness of MIR lies in its ability to uncover intricate relationships between features and the target variable. It excels in scenarios where traditional linear methods fall short. Researchers and data scientists have found that MIR often leads to improved predictive models, as it captures essential non-linear dependencies that would otherwise remain hidden.

In summary, Mutual Information Regression is a powerful feature selection technique that excels in non-linear, complex data relationships. Its ability to measure the information gain between features and the target variable, along with its inherent advantages, makes it a valuable asset in the data scientist's toolkit. By using MIR, one can unlock the potential of their data, leading to more accurate and insightful predictive models across a wide range of applications.

3.2 Feature Extraction - Principal Component Analysis (PCA)

Feature extraction plays a pivotal role in data analysis, where the goal is to uncover the most salient patterns within a dataset while reducing its dimensionality. Principal Component Analysis (PCA)[21] stands as one of the most prominent techniques in this realm. In this writeup, we delve into the workings of PCA, its primary purpose, and the array of advantages it offers, particularly its ability to remove correlation and produce independent features.

3.2.1 Principal Component Analysis (PCA): An Overview

PCA[21] is a dimensionality reduction technique that allows us to transform a high-dimensional dataset into a lower-dimensional one while retaining as much of the original information as possible. The central idea behind PCA is to project the data onto a new coordinate system where the axes (principal components) are orthogonal and capture the most variance.

3.2.2 The Working Mechanism of PCA

- **Decomposition:** PCA starts by decomposing the dataset into its principal components, which are linear combinations of the original features. These components are chosen to maximize the variance explained.
- **Variance Maximization:** The first principal component captures the maximum variance in the data. Subsequent components are chosen in a way that they are orthogonal to the previous ones and maximize the remaining variance..
- **Dimension Reduction:** By retaining a subset of the principal components, PCA effectively reduces the dimensionality of the data. It is particularly useful when dealing with high-dimensional datasets or datasets with correlated features.

3.2.3 Advantages of PCA

PCA offers a multitude of advantages that make it a powerful tool for feature extraction and dimensionality reduction:

- **Correlation Removal:** PCA excels at identifying and removing correlations between features. This is particularly useful when working with datasets in which features are interrelated.
- **Independence:** The principal components produced by PCA are orthogonal, meaning they are independent of each other. This independence is invaluable for feature extraction, as it ensures that the selected features do not carry redundant information.
- **Dimension Reduction:** PCA's ability to reduce the dimensionality of a dataset is a critical advantage. It simplifies the data representation, making it more manageable for modeling and analysis.
- **Variance Retention:** While reducing dimensionality, PCA strives to retain as much variance as possible. This means that important patterns and structures within the data are preserved, even with a lower number of features.
- **Noise Reduction:** PCA can filter out noise and capture the underlying signal within a dataset, leading to more robust and accurate models.

3.2.4 Effectiveness of Principal Component Analysis

The effectiveness of PCA is evident in its widespread use across various domains, from image and signal processing to finance and healthcare. By removing correlations, extracting independent features, and reducing dimensionality, PCA not only simplifies the data but also enhances the interpretability and predictive power of models.

In conclusion, Principal Component Analysis is a versatile and robust technique for feature extraction and dimensionality reduction. Its ability to remove correlations,

produce independent features, and reduce dimensionality while preserving essential information makes it an indispensable tool for data scientists and researchers. Whether in the pursuit of clearer insights, more efficient modeling, or improved predictive accuracy, PCA remains a powerful ally in the world of data analysis.

3.3 Neural Networks and Deep Learning in Time Series Regression Prediction

Time series data, with its sequential nature and temporal dependencies, presents a unique challenge for regression prediction. In recent years, the advent of neural networks and deep learning models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, has revolutionized the way we approach time series regression problems. In this writeup, we explore how these sophisticated models have redefined time series prediction, their advantages, and the immense potential they offer.

3.3.1 Neural Networks in Time Series Regression

Neural networks, inspired by the human brain, have demonstrated remarkable capabilities in capturing complex patterns within sequential data. In time series regression, feedforward neural networks, with their ability to model non-linear relationships, have paved the way for more accurate predictions. Deep neural networks, equipped with multiple layers, are adept at hierarchically learning representations of the data.

3.3.1.1 Advantages of Neural Networks

- **Non-linearity:** Neural networks are not bound by the assumptions of linearity, making them highly effective at capturing non-linear relationships present in time series data.
- **Temporal Dependencies:** Recurrent neural networks, such as LSTMs, inherently account for temporal dependencies in data. This makes them well-suited

for sequential data analysis, as they can remember and utilize past information in predictions.

- **Feature Learning:** Deep neural networks automatically learn relevant features from the data, reducing the need for manual feature engineering.

3.3.2 Recurrent Neural Networks (RNNs) in Time Series Regression

RNNs[23] are a class of neural networks that are specifically designed to handle sequential data. They maintain an internal state that evolves over time, allowing them to consider past inputs while making predictions for the future. In time series regression, RNNs excel at modeling dynamic patterns and dependencies.

3.3.2.1 Long Short-Term Memory (LSTM) Networks in Time Series Regression

LSTMs[20] are a specialized type of RNN designed to mitigate the vanishing gradient problem, which often hinders training in traditional RNNs. LSTMs have proven to be highly effective in modeling long-range dependencies and capturing subtle nuances in time series data.

3.3.2.2 Advantages of RNNs and LSTMs

- **Temporal Modeling:** RNNs and LSTMs are explicitly designed for sequential data, allowing them to capture complex temporal dependencies that may be challenging for other models.

- **Long-Term Memory:** LSTMs, in particular, excel at retaining long-term memory, making them effective at capturing patterns with extended temporal dependencies.

- **Variable Sequence Length:** RNNs and LSTMs can handle variable-length sequences, offering flexibility in real-world applications where data collection

may not adhere to a fixed schedule.

3.3.3 Effectiveness of Deep Learning in Time Series Regression

The application of deep learning models like RNNs[23] and LSTMs[20] to time series regression has yielded substantial improvements in predictive accuracy. These models are particularly valuable when dealing with data exhibiting complex, non-linear patterns and temporal dependencies. Their ability to automatically learn relevant features and capture intricate relationships within sequential data offers a significant advantage in a wide range of domains, from finance and healthcare to climate forecasting and beyond.

In conclusion, neural networks and deep learning models, particularly RNNs and LSTMs, have redefined the landscape of time series regression prediction. Their non-linearity, capacity to model temporal dependencies, and automatic feature learning have made them invaluable tools for data scientists and researchers. By embracing these advanced techniques, we unlock the potential to make more accurate, insightful, and forward-looking predictions in the realm of time series regression.

3.4 Feedforward Neural Networks as Encoders for Dimensionality Reduction in Time Series Regression

In the ever-evolving landscape of predictive modeling, dimensionality reduction is a key component in simplifying complex datasets while preserving their inherent information. Feedforward neural networks, a class of artificial neural networks, can be instrumental in this task by serving as encoders. In this writeup, we explore how feedforward neural networks can be employed to encode features into lower dimensions and how these encodings can empower deep learning models, such as RNNs and

LSTMs, to tackle time series regression prediction with precision and efficiency.

3.4.1 Feedforward Neural Networks as Encoders

Feedforward neural networks, also known as multilayer perceptrons, are designed to model complex relationships within data. When used as encoders, they transform high-dimensional feature vectors into compact, lower-dimensional representations. The network architecture typically includes an input layer, one or more hidden layers, and an output layer. During the encoding process, the network learns to capture and retain the most relevant features, effectively reducing the dimensionality of the data.

3.4.1.1 Advantages of Feedforward Neural Networks as Encoders:

- **Non-linearity:** These networks are adept at modeling non-linear relationships, making them suitable for encoding complex patterns within feature sets.
- **Automatic Feature Selection:** Feedforward neural networks automatically perform feature selection, as they learn which attributes are most informative for the task at hand.
- **Generalization:** The encoded representations often generalize well, making them effective for capturing essential information while reducing noise and redundancy.

3.4.2 Empowering Deep Learning Models with Encodings

Once the feedforward neural network has encoded the high-dimensional feature vectors, the resulting lower-dimensional representations can be passed as input to deep learning models designed for time series regression, such as RNNs and LSTMs. These deep learning models are capable of understanding temporal dependencies and complex patterns, allowing them to make accurate predictions based on the encodings.

3.4.3 Effectiveness of the Combined Approach

The synergy between feedforward neural networks as encoders and deep learning models like RNNs and LSTMs is a powerful one. It allows for feature reduction while preserving the essential aspects of the data. By encoding features into lower dimensions, the models become more efficient and capable of capturing intricate temporal dependencies, resulting in accurate time series regression predictions.

In conclusion, feedforward neural networks, when utilized as encoders, offer a means to reduce feature dimensionality and enhance the efficiency of deep learning models. By leveraging this combined approach, we unlock the potential for more accurate and insightful time series regression predictions, with applications spanning various domains, from finance to healthcare and beyond.

3.5 Prophet Model: A Forecasting Marvel for Time Series Regression

Prophet[36] is an open-source forecasting tool designed to handle time series data with daily observations that display patterns on multiple time scales. It was specifically engineered to address the challenges that arise when dealing with time series regression, such as holidays, seasonality, and abrupt changes in trends. Prophet employs a decomposable time series model that accounts for these diverse components.

3.5.1 The Working Mechanism of Prophet

- **Seasonality and Holidays:** Prophet recognizes both yearly and weekly seasonality. It allows the inclusion of holidays and special events, acknowledging their impact on the data.
- **Trend Components:** Prophet decomposes the data into three key components—trend, seasonality, and holiday effects. The trend captures the underly-

ing trajectory of the time series.

- **Flexibility:** Prophet is highly adaptable and can handle missing data points, outliers, and abrupt changes in trends. It does not require manual data preprocessing.
- **Automatic Changepoint Detection:** The model automatically detects changepoints, where the time series' trajectory shifts significantly.

3.5.2 Advantages of Prophet in Time Series Regression

- **Ease of Use:** Prophet is known for its simplicity and user-friendly design. It allows analysts and data scientists to work efficiently without an extensive background in time series forecasting.
- **Holiday Effects:** The model's ability to account for holidays and special events is particularly advantageous in various domains, from retail to healthcare.
- **Transparency:** Prophet offers transparency in forecasting, as it decomposes the time series into interpretable components. This makes it easy to understand the model's predictions.
- **Automatic Component Selection:** The model automatically selects the relevant components (trend, seasonality, holidays), reducing the need for manual feature engineering.

3.5.3 Applications of Prophet Model

The versatility of the Prophet model extends across numerous domains, including but not limited to sales forecasting, demand planning, financial market predictions, and epidemiological modeling. Its intuitive interface, robust performance, and adaptability to different data scenarios make it a valuable tool for data analysts and businesses seeking reliable time series regression predictions.

In conclusion, the Prophet model has established itself as a forecasting marvel in the world of time series regression. Its ease of use, transparency, and adaptability to diverse data scenarios have made it an invaluable asset for predictive modeling. Whether predicting sales trends, disease outbreaks, or financial market movements, Prophet has proven its mettle as an accurate and versatile forecasting tool.

3.6 DeepAREstimator: A Unified Solution for Multi-Time Series Regression

DeepAREstimator[29], based on the DeepAR (Deep Autoregressive) architecture, is designed to handle a multitude of time series together. It leverages a deep neural network to capture complex temporal dependencies and patterns across the time series data. The model excels in making accurate predictions and generating probabilistic forecasts for multiple time series simultaneously.

3.6.1 Key Features of DeepAREstimator

- **Shared Knowledge:** DeepAREstimator capitalizes on the shared information across the multiple time series. By training a single model, it harnesses the commonalities and dissimilarities among the time series to enhance forecasting accuracy.
- **Autoregressive Structure:** The model employs an autoregressive structure that can model the time dependencies in each series effectively. This allows it to understand how each series influences its future values.
- **Probabilistic Forecasting:** DeepAREstimator provides probabilistic forecasts, offering a range of possible outcomes. This is essential for risk assessment and uncertainty management in real-world applications.
- **Scalability:** The model is highly scalable, making it suitable for a wide range of use cases, from forecasting sales data for various products to predicting energy

consumption across different regions.

3.6.2 Effectiveness of DeepAREstimator in Multi-Time Series Regression

The DeepAREstimator model has proven to be remarkably effective in making predictions across multiple time series. It capitalizes on the inherent similarities and dependencies shared among the time series, resulting in more accurate and coherent forecasts. This not only simplifies the modeling process but also ensures that insights are extracted comprehensively from all the data.

In conclusion, the DeepAREstimator model is a powerful and versatile tool for multi-time series regression prediction. Its ability to handle multiple time series together while producing probabilistic forecasts makes it an invaluable asset in diverse fields. Whether forecasting sales for a range of products or predicting energy consumption across various locations, DeepAREstimator has demonstrated its effectiveness in simplifying the modeling process and generating accurate, comprehensive predictions.

CHAPTER 4

Building A New Novel COVID-19 Dataset

In response to the limitations of existing COVID-19 datasets, a novel and meticulously curated dataset has been painstakingly constructed. This new dataset has been developed with a keen awareness of the shortcomings encountered in previous datasets, aiming to overcome the challenges posed by limited feature sets, data quality issues, and the absence of crucial contextual information. By addressing these limitations, the novel COVID-19 dataset aspires to empower researchers and analysts with a more comprehensive and reliable resource, capable of facilitating deeper insights and more accurate predictions related to the COVID-19 pandemic.

4.1 Data Collection and Description

In our persistent endeavor to comprehend the profound impact of COVID-19 on Canada, we have undertaken the compilation of a comprehensive and meticulously curated dataset that spans all ten provinces: Alberta (AB), British Columbia (BC), Manitoba (MB), New Brunswick (NB), Newfoundland and Labrador (NL), Nova Scotia (NS), Ontario (ON), Prince Edward Island (PE), Quebec (QC), Saskatchewan (SK), and the three territories: Northwest Territories (NT), Nunavut (NU), and Yukon (YT). This extensive dataset, as presented in Tables 3.1.1 and 3.1.2, serves as a foundational resource for the detailed analysis of the pandemic's multifaceted effects on various aspects of life in Canada.

Our dataset is characterized by its encompassing nature, offering vital insights into an array of factors that collectively contribute to the understanding of the COVID-19 pandemic. These factors include Environmental Factors, Government Actions, Medical Factors, Mobility, Weather Conditions, Age Distribution, Geographical and Land Distribution, Economics, Health Characteristics, Diseases, and societal factors like Bad Habits in the population. The information we provide seeks to unravel the intricate interplay of these variables and their impact on the pandemic's progression in Canada.

Data collection for this extensive dataset commenced in early 2020, with a primary focus on capturing the initial COVID-19 Total Daily Cases, Total Daily Dealths, Total Daily Hospitalisations, Daily Cases, Daily Dealths, Daily Hospitalisations in each province and territory. This rigorous data collection process continued uninterrupted until mid-September 2022, thereby encapsulating a substantial time frame of the pandemic's evolution in the country. The dataset offers a panoramic view of the pandemic's progression in Canada, capturing pivotal moments, shifts, and trends.

Table 4.1.1: NON-MEDICAL FEATURES FOR DATA COLLECTION

Feature Group	**Feature**
Date	Date
Region	Region
Government Actions	Action-Provincial-1 to 22; Intervention Category-Provincial-1 to 22; Intervention Type-Provincial-1 to 22; Action-Country-1 to 4; Intervention Category-Country- 1 to 4; Intervention Type-Country-1 to 4

Mobility	retail and recreation percent change from baseline region; grocery and pharmacy percent change from baseline region; parks percent change from baseline region; transit stations percent change from baseline region; workplaces percent change from baseline region; residential percent change from baseline region; retail and recreation percent change from baseline country; grocery and pharmacy percent change from baseline country; parks percent change from baseline country; transit stations percent change from baseline country; workplaces percent change from baseline country; residential percent change from baseline country; Holiday; Domestic movements; Transborder movements; Other international movements; International travellers entering or returning to Canada
Weather Conditions	Mean Temp (°C); Total Precip (mm); SpeedOfWind(km/h)

Age Distribution	Males-0 to 4 years; Males-5 to 9 years; Males-10 to 14 years; Males-15 to 19 years; Males-20 to 24 years; Males- 25-29 years; Males-30-34 years; Males-35-39 years; Males- 40-44 years; Males-45 to 49 years; Males-50 to 54 year; Males-55 to 59 years; Males-60 to 64 years; Males-65- 69 years; Males-70-74 years; Males-75-79 years; Males- 80 to 84 years; Males-85 to 89 years; Males-90-94 years; Males-95-99 years; Males-100 years and over; Females-0 to 4 years; Females-5 to 9 years; Females-10 to 14 years; Females-15 to 19 years; Females-20 to 24 years,Females- 25-29 years; Females-30-34 years; Females-35-39 years; Females-40-44 years; Females-45 to 49 years; Females- 50 to 54 years; Females-55 to 59 years; Females-60 to 64 years; Females-65-69 years; Females-70-74 years; Females- 75-79 years,Females-80 to 84 years; Females-85 to 89 years; Females-90-94 years; Females-95-99 years; Females- 100 years and over; Total-All Ages-Male; Median Age- Male; Average Age-Male; Total-All Ages-Female; Median Age-Female; Average Age-Female
Geographical and land Distribution	Total private dwellings, 2021; Private dwellings occupied by usual residents, 2021; Land area in square kilometres, 2021; Population density per square kilometre, 2021
Economical	Gross domestic product (GDP) at basic prices; Life expectancy (in years) at age 0; Food insecure, moderate or severe Percentage of Males; Food insecure, moderate or severe Percentage of Females; Unemployment rate Males; Participation rate Males; Employment rate Males; Unemployment rate Females; Participation rate Females; Employment rate Females; Human Development Index 2019.

The data included in our dataset have been meticulously gathered from diverse

sources, including Statistics Canada (StatsCan) [33], the Canadian Institute for Health Information (CIHI) [13], Google Mobility [18], GitHub repositories [15] and various other data repositories [32], [37] among others. These diverse sources collectively contribute to the richness and comprehensiveness of the dataset, enabling a multifaceted analysis of the COVID-19 pandemic in Canada.

Table 4.1.2: MEDICAL FEATURES FOR DATA COLLECTION

Feature Group	**Feature**
Medical Factors	Daily Tests Completed; Total ICU occupancy; No of Hospitals in Province 2021
Vaccination Coverage	Vaccine Coverage Percent Dose 1 to 4; Vaccine administration Total Doses; Vaccine administration Dose 1 to 3

Health Characteristics of Population	Perceived health, very good or excellent-Male; Perceived health, fair or poor-Male; Perceived mental health, very good or excellent-Male; Perceived mental health, fair or poor-Male; Perceived life stress, most days quite a bit or extremely stressful-Male; Body mass index, adjusted self-reported, adult (18 years and over), overweight-Male; Body mass index, adjusted self-reported, adult (18 years and over), obese-Male; Body mass index, self-reported, youth (12 to 17 years old), overweight or obese-Male; Arthritis-Male; Diabetes-Male; Asthma-Male; Chronic obstructive pulmonary disease-Male; High blood pressure- Male; Mood disorder-Male; Self-reported physical activity, 150 minutes per week, adult-Female; Self-reported physical activity, average 60 minutes per day, youth-Female; Breast milk feeding initiation-Female; Exclusive breastfeeding, at least 6 months-Female; Fruit and vegetable consumption- Female; Sense of belonging to local community, somewhat strong or very strong-Female; Life satisfaction, satisfied or very satisfied-Female; Percentage of persons in low income-Male; Percentage of persons in low income-Female; Percentage of persons with unmet health care needs-Male(2020); Percentage of persons with unmet health care needs-Female(2020).

Diseases and Bad Habits in Population	Current smoker, daily or occasional-Male; Current smoker, daily-Male; Cannabis use-Male; Cannabis frequency of use in the past months, daily or almost daily-Male; Heavy drinking-Male; Selfreported physical activity, 150 minutes per week, adult- Male; Self-reported physical activity, average 60 minutes per day, youth-Male; Fruit and vegetable consumption- Male; Sense of belonging to local community, somewhat strong or very strong-Male; Life satisfaction, satisfied or very satisfied-Male; Has a regular healthcare provider-Male; Influenza immunization in the past 12 months-Male; Perceived health, very good or excellent-Female; Perceived health, fair or poor-Female; Perceived mental health, very good or excellent-Female; Perceived mental health, fair or poor-Female; Perceived life stress, most days quite a bit or extremely stressful-Female; Body mass index, adjusted self-reported, adult (18 years and over), overweight- Female; Body mass index, adjusted self-reported, adult (18 years and over), obese-Female; Body mass index, self-reported, youth (12 to 17 years old), overweight or obese-Female; Arthritis-Female; Diabetes-Female; Asthma- Female; Chronic obstructive pulmonary disease-Female; High blood pressure-Female; Mood disorder-Female; Current smoker, daily or occasional-Female; Current smoker, daily-Female; Cannabis use-Female; Cannabis frequency of use in the past months, daily or almost daily-Female; Heavy drinking-Female; Has a regular healthcare provider- Female; Influenza immunization in the past 12 months- Female.

Through our comprehensive framework, our ultimate objective is to harness the power of data and cutting-edge predictive techniques to forecast a range of critical target values. These values encompass the broader spectrum of the COVID-19 pandemic's impact, spanning across multiple dimensions. Specifically, our framework

is meticulously designed to predict essential indicators such as Total Daily Cases, which serve as a key metric in understanding the overall progression of the virus. Furthermore, we aim to provide insights into the Total Daily Deaths, shedding light on the severity of the pandemic's toll on human lives. Additionally, our predictive capabilities extend to Total Daily Hospitalizations, a critical parameter that reflects the strain on healthcare infrastructure. However, our commitment to precision and timely insights doesn't stop there; we delve into the realm of daily dynamics, offering predictions for Daily Cases, Daily Deaths, and Daily Hospitalizations. By encompassing this comprehensive array of target values, our framework equips stakeholders with the invaluable ability to anticipate and respond effectively to the evolving landscape of the COVID-19 pandemic.

Table 4.1.3: Target Values Table

Target Values/ Prediction Values
Total Daily Cases
Total Daily Dealths
Total Daily Hospitalisations
Daily Cases
Daily Dealths
Daily Hospitalisations

This expansive window of data collection encompasses the critical phases of the COVID-19 pandemic, amounting to a total of 11,995 rows, with each row representing a distinct day in this timeline. Within this dataset, a multitude of columns stands as a testament to the breadth of information we've diligently gathered. These columns are laden with a rich tapestry of features, each meticulously recorded on a daily basis. This multifaceted approach allows us to encapsulate the dynamic nature of the pandemic's progression, unveiling a wealth of insights into its various facets. From epidemiological statistics to environmental factors, government actions, and more, our dataset provides a holistic view of the evolving COVID-19 landscape, enabling

comprehensive analyses and informed decision-making.

4.2 Missing Values/Null Values Treatment

In our relentless pursuit of data accuracy and completeness, we encountered a minor gap in our dataset: the absence of continuous values for certain features. To address this gap, we employed a systematic and data-driven approach, harnessing the power of time-based interpolation. This technique enabled us to seamlessly infill the missing continuous values, ensuring the dataset's integrity and reliability. By considering the temporal context and leveraging data from related time points, we were able to derive precise estimations for the missing continuous features, taking into account the unique characteristics of each region. Through this interpolation process, we not only bridged the data gaps but also maintained the temporal coherence of the dataset, guaranteeing that our analytical endeavors are underpinned by a comprehensive and robust foundation of information.

4.3 Data Versioning for Experimentation And Best Performaning Dataset

Our meticulous data preprocessing efforts led to the creation of three distinct versions of our database, each tailored to address the intricate nature of categorical features and their one-hot encoding. In the first iteration, aptly named **Version 1 (Transfer Learning Dataset)**, we tackled the challenge by one-hot encoding all categorical features, with the exception of Region and Provincial Action and Intervention Category Group Features. For the former, we encoded the top 5 values, while the remaining categorical attributes were transformed for all unique values. This approach resulted in a database comprising 578 features.

In **Version 2**, we refined our strategy by expanding one-hot encoding to include all categorical features, even encompassing the Region attribute. Here, the top 5 values for each feature were considered, resulting in a database with 438 features.

The evolution of our data manipulation journey culminated in **Version 3**, a database of 445 features. In this iteration, the Region feature was one-hot encoded for all of its values, while the other categorical attributes were one-hot encoded for their respective top 5 values.

Table 4.3.1: Dataset Comparison Table

Parameter/ Studies	**Previous Research Study[25]**	**Motivational Research Study[26]**	**Our Research Study[31]**
Dataset	John Hopkins Dataset[16]	Custom Dataset	Custom Dataset
Feature Group	0	4	12
Total Versions	1	1	3
Total Features other than Target Variables	0	13	445
Data Level	Country	Province	Province

Intriguingly, our empirical analyses revealed that the utilization of **Version 3** consistently yielded the most promising results among the three iterations. This noteworthy outcome underscores a compelling observation: **surpassing the region's inclusion and expanding the feature set did not necessarily translate into superior predictive performance.** Instead, it emphasizes the importance of targeted, data-driven decisions in feature engineering and the nuanced interplay between feature richness and model effectiveness.

CHAPTER 5

Methodology

In this chapter, we present a unified framework that bridges the realms of High Cardinality Multivariate Multi Time Series (HCMVMT) datasets and the prediction of COVID-19 epidemic trends. The pressing need for a comprehensive approach that seamlessly integrates these domains becomes evident as we delve into the intricacies of framework architecture. We dissect the framework into two key dimensions: one designed to address the prediction problem as a regression task and another tailored for time series forecasting. This unified framework not only offers a robust solution to the challenges posed by both domains but also fosters interdisciplinary insights, paving the way for more accurate and insightful predictions.

5.1 A Unified Framework for HCMVMT Datasets and COVID-19 Epidemic Trend Prediction

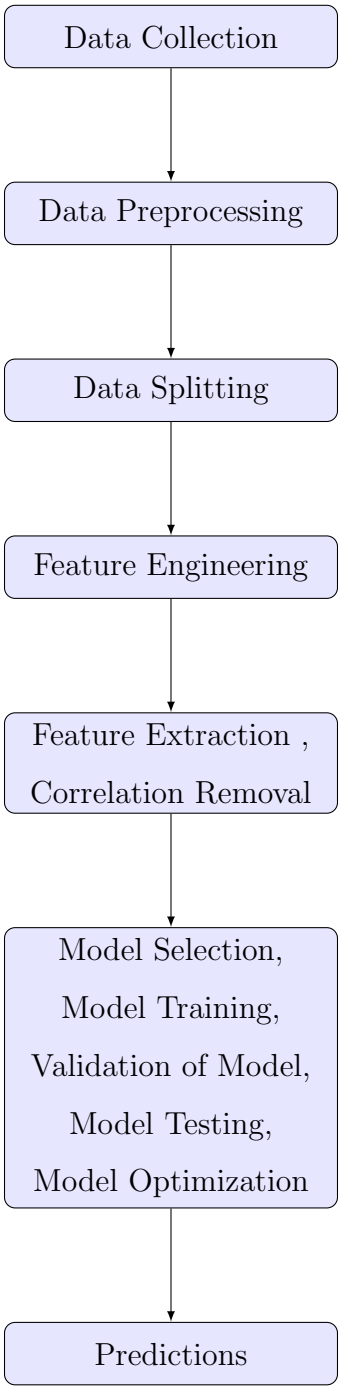
The present research embarks on a comprehensive journey into the heart of its methodology—a dynamic and adaptable framework designed to provide holistic solutions for the intricate challenges posed by High Cardinality Multi-Variate Multi-Time Series (HCMVMT) datasets. This framework transcends its immediate applications to extend a helping hand to the realm of COVID-19 epidemic trend prediction, unveiling a versatile tool capable of illuminating the complex dynamics that underlie a global health crisis. In this section, we delve into the architecture, components, and strategies that collectively shape this transformative methodology.

5.2 The Need for a Unified Framework

The contemporary landscape of data analytics and predictive modeling presents an intriguing paradox. On one hand, there is a proliferation of high-dimensional, multi-variate, multi-time series datasets, representing diverse domains from finance to healthcare. On the other hand, the COVID-19 pandemic has ignited an urgent demand for accurate, real-time predictions of epidemic trends, demanding the assimilation of data from multiple sources and the incorporation of varying data structures. The challenge before us is to not only unravel the intricacies of HCMVMT datasets but to also cater to the pressing need for precise COVID-19 epidemic trend forecasting. It is within this context that the unified framework takes center stage.

5.3 Framework Architecture for Solving as a Regression Problem

Proposed Hybrid Regression Framework Flowchart



- **Data Collection:** Load the data from the new novel custom dataset built from various sources.
- **Data Preprocessing:**
  - Address null values in features through time-based interpolation for continuous features within the same region.
  - Perform one-hot encoding of essential categorical features, the method of which is version-dependent.
  - Normalize all the features to ensure consistent scaling and prevent bias in model training.
- **Data Splitting:** Divide the dataset into distinct sets for training (70%), validation (10%), and testing (20%). This partitioning facilitates the robust evaluation of the developed models.
- **Feature Engineering:** Utilize the Mutual Information Regression (MIR) method to identify features that exhibit a strong correlation with the target values. This step is pivotal for selecting the most informative attributes.
- **Feature Extraction and Correlation Removal:**To mitigate issues arising from high correlation among selected features, apply Principal Component Analysis (PCA). PCA transforms the dataset into new, independent features, reducing dimensionality and enhancing model performance.
- **Model Selection and Development:** Following feature extraction and Correlation Removal, engage in multiple iterations to select or construct the best-performing deep learning models for regression. Consider a variety of architectures, including feed forward neural networks[30] and Random Forest.
- **Model Training:** Train the selected models using the training dataset. Implement state-of-the-art training techniques and optimization algorithms to ensure convergence and the generation of reliable models.

- **Model Validation:** Employ the validation dataset to fine-tune hyperparameters and assess the model's performance. Evaluate the models against various metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2).
- **Model Testing:** Verify the models' generalization performance using the dedicated testing dataset. This step is essential for gauging how well the models will perform on new, unseen data.
- **Model Optimization:** Continuously optimize the models based on the performance indicators and insights derived from validation and testing phases. Make necessary adjustments to enhance predictive accuracy.

This comprehensive framework is designed to address the complexities of the regression problem while ensuring that the developed models can accurately predict target values such as Total Daily Cases, Total Daily Deaths, Total Daily Hospitalizations, Daily Cases, Daily Deaths, and Daily Hospitalizations. The iterative nature of the framework allows for continuous improvement, making it adaptable to evolving data and research needs.

5.4 Regression Model Performance Variability in Predicting COVID-19 Epidemic Trends

Our Deep Neural Network (DNN) based regression model framework has demonstrated considerable success in predicting COVID-19 epidemic trends, particularly when forecasting Total Daily Cases, Total Daily Deaths, and Total Daily Hospitalizations. **The incorporation of region-based features has played a pivotal role in setting up a reliable baseline, allowing our framework to generate accurate predictions in these cases.** The performance of our model under these circumstances has been commendable, aligning well with real-world data.

However, the scenario changes when we shift our focus to predicting the change in Daily Cases, Daily Deaths, and Daily Hospitalizations. **It becomes evident that**

the baseline established by region features is no longer a strong predictor of these dynamic variables. The distribution of such changes can exhibit similarities across multiple regions, making it challenging to discern patterns and trends solely based on regional data.

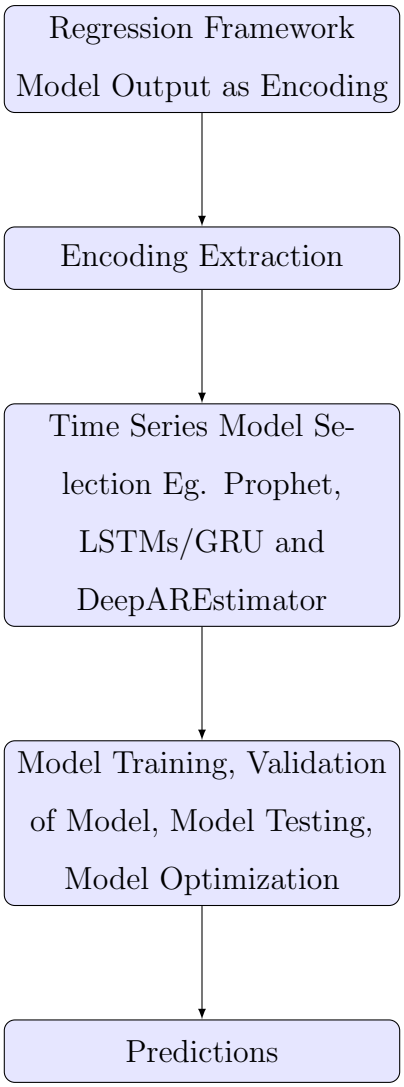
One of the primary limitations we encounter is our model's inability to capture the temporal dependencies of previous dates and their associated features when making predictions for the current date. In the context of predicting changes in Daily Cases, Daily Deaths, and Daily Hospitalizations, understanding the historical context and its impact on the present is crucial. The lack of mechanisms to account for such dependencies restricts the predictive capacity of our regression framework.

As a result of these limitations, it becomes evident that a different approach is required to address the prediction of changes in Daily Cases, Daily Deaths, and Daily Hospitalizations. While our current framework excels in forecasting cumulative numbers, the intricate dynamics involved in daily changes necessitate the application of time series modeling techniques. Time series frameworks are designed to handle dependencies over time and are better suited to capture the nuances and fluctuations in the data. Therefore, it is imperative to develop and implement a dedicated time series framework that can effectively predict the changes in these critical COVID-19 statistics.

In the subsequent section, we will explore the development and integration of time series models into our framework to address these limitations and enhance our predictive capabilities for daily changes in COVID-19 epidemic trends.

5.5 Framework Architecture for Solving as a Time Series

Proposed Deep Learning Time Series Framework



To address the limitations encountered in our regression model framework, we propose the integration of a dedicated time series framework. This new approach is specifically designed to predict changes in Daily Cases, Daily Deaths, and Daily Hospitalizations, where the dynamics differ significantly from forecasting cumulative numbers. The proposed framework is a multi-step process aimed at optimizing the prediction accuracy and reliability.

- **Input: Utilizing Neural Network Encoding**

  The input to our time series framework is based on the encoding derived from the last layer of the neural network in our regression model framework. This encoding encapsulates the predictive information extracted from the regression model's output for Total Daily Cases, Total Daily Deaths, and Total Daily Hospitalizations.

  The encoding acts as a condensed representation of the baseline predictions generated by the regression framework. It captures the essential features and trends observed in the cumulative statistics. This encoding is used as a starting point for the time series models, facilitating a seamless transition from regression to time series forecasting.

  By incorporating this neural network encoding, we ensure that the relevant information and patterns obtained from the regression model are leveraged in our time series framework, enhancing the accuracy and reliability of predictions for changes in Daily Cases, Daily Deaths, and Daily Hospitalizations.

- **Encoding Extraction**

  Before proceeding with time series modeling, we employ encoding techniques to distill the wealth of information generated by the regression framework. These encodings encapsulate the essential features and patterns observed in the cumulative statistics. The encoding process prepares the data for further analysis by capturing the significant characteristics of the dataset.

- **Time Series Model Selection**

  The heart of our time series framework lies in the selection of appropriate models for predicting changes in Daily Cases, Daily Deaths, and Daily Hospitalizations. We explore several time series models known for their effectiveness in capturing temporal dependencies and dynamic patterns. The primary models under consideration include:

- **Prophet:** Prophet is a robust forecasting tool that excels in capturing daily and weekly seasonality, as well as holidays and sudden changes in trends.
- **LSTMs/GRU:** Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are deep learning models designed for sequence prediction. These models can effectively capture complex temporal dependencies.
- **DeepAREstimator:** The DeepAREstimator model, known for its adaptability and scalability, is particularly suitable for multi-variate, multi-time series datasets. It offers a versatile approach to capturing intricate dynamics.

The choice of model depends on the specific characteristics of the data and the nature of the predictions required. We aim to identify the model that best suits the complexity of the COVID-19 dataset.

- **Model Training, Validation, Testing, and Optimization**

  Once the time series models are selected, we initiate the training phase. The models are trained on the encoded dataset, where they learn to capture temporal dependencies and patterns. After training, we move on to the validation stage, fine-tuning hyperparameters and assessing the models' performance against various evaluation metrics.

  The testing phase follows, where the models are evaluated rigorously to ensure their capability to predict changes in Daily Cases, Daily Deaths, and Daily Hospitalizations. This comprehensive testing process helps us gauge the accuracy and reliability of the predictions generated by each model.

  To further enhance the predictive capabilities of the models, we undertake an optimization phase, fine-tuning parameters and making necessary adjustments based on the evaluation results.

- **Predictions**

The final output of our time series framework is a set of predictions for changes in Daily Cases, Daily Deaths, and Daily Hospitalizations. These predictions are generated by the selected time series models and are critical for understanding the dynamic nature of the COVID-19 epidemic trends. The predictions provide valuable insights into the fluctuating patterns and temporal dependencies, enabling informed decision-making and proactive response strategies.

In the subsequent chapters, we will delve into the details of each stage of our time series framework, exploring the model selection, training, validation, testing, optimization, and the insights gleaned from the predictions. This comprehensive approach aims to overcome the limitations of our regression framework and provide a robust solution for predicting changes in COVID-19 epidemic trends.

CHAPTER 6

Experiments and Results

In the chapter, we embark on an exhaustive exploration of the intricate experimentation setup and evaluation framework that underpin our research. This chapter serves as the fulcrum of our investigative efforts, where we meticulously delineate the methodologies, configurations, and rigorous evaluation metrics harnessed to scrutinize the performance of our unified framework. With meticulous precision, we navigate the experimental landscape, offering comprehensive insights into each trial conducted, the diverse data scenarios examined, and the variations in model deployments explored. These rigorous experiments culminate in the unveiling of results that validate our framework's prowess in predicting COVID-19 epidemic trends and tackling the multifaceted challenges posed by High Cardinality Multivariate Multi Time Series (HCMVMT) datasets. Join us on this journey of meticulous experimentation and discovery, where empirical evidence illuminates the transformative potential of data-driven decision-making in these complex domains. Embark on this comprehensive exploration of meticulous experimentation and discovery, where empirical evidence unveils the profound potential of data-driven decision-making within these multifaceted domains.

6.1 Experimentation Setup for Regression Framework

In this section, we elaborate on the meticulous setup of our experiments for the regression framework, aimed at predicting the numbers of Total Daily Cases, Total Daily

Deaths, and Total Daily Hospitalizations. The setup encompasses a range of input data, algorithms/models employed, and the environment in which the experiments were conducted.

6.1.1 Input: New Dataset

The cornerstone of our experimentation lies in the utilization of a new and comprehensive dataset. This dataset acts as the bedrock upon which we conduct our predictive analyses, encompassing a diverse array of features and temporal dependencies.

6.1.2 Algorithms/Models

To assess the performance of our regression framework, we explore multiple algorithms and models:

- **K-Nearest Neighbors (KNN)[22]**: A traditional machine learning algorithm for regression tasks.
- **Random Forest**: An ensemble learning method that leverages decision trees for regression.
- **Deep Neural Network[30]**: A fundamental deep learning model for regression.
- **Deep Neural Network with Regression Framework**: Our proprietary deep learning architecture, designed to handle High Cardinality Multivariate Multi Time Series datasets.

These models are meticulously crafted and fine-tuned to address the complex task of predicting COVID-19 epidemic trends.

6.1.3 Training Environment

All experimental studies were meticulously conducted within a controlled environment to ensure the reproducibility and accuracy of our results. The training environment

was established as follows:

- **Operating System**: A 64-bit Debian GNU/Linux 9.11 operating system provided the foundation for our experiments.
- **Hardware**: The experimental environment was powered by an Intel (R) Xeon (R) Gold CPU @ 2.20GHz, complemented by 16 GB of RAM.
- **GPU Acceleration**: To enhance deep learning model training, we leveraged NVIDIA Tesla K80, boasting 12GB of GPU memory.
- **Deep Learning Framework**: The PyTorch deep learning framework was employed to harness the potential of neural networks in our models.

This carefully orchestrated environment ensured that our experiments were conducted under consistent and controlled conditions.

6.1.4 Output

The output of our experiments constitutes the core performance metrics that gauge the effectiveness of our regression framework. The following output metrics were assessed:

- **Total Daily Cases**
- **Total Daily Deaths**
- **Total Daily Hospitalizations**

These metrics are the yardstick against which we evaluate the predictive capabilities of our models and framework.

In the subsequent sections, we delve into the results and insights derived from these meticulously designed experiments, shedding light on the efficacy of our regression framework in the domain of COVID-19 epidemic trend prediction."

6.2 Experimentation Setup for Deep learning Time Series Framework

In this section, we provide a comprehensive insight into the meticulous setup of our experiments for the Time Series Framework, tailored for predicting Daily Cases, Daily Deaths, and Daily Hospitalizations. The setup encompasses the input data, a range of algorithms and models employed, and the environment in which the experiments were meticulously conducted.

6.2.1 Input: Encodings from Regression Framework

The foundational input for our Time Series Framework is the encodings derived from the Regression Framework's deep neural network model. These encodings, which encapsulate the essence of the COVID-19 pandemic trends, serve as a bridge between the two frameworks, enabling a seamless transition from regression to time series analysis.

6.2.2 Algorithms/Models

To evaluate the capabilities of our Time Series Framework, we explore a spectrum of algorithms and models, each tailored to handle the nuances of time series data:

- **ARIMA[19]/Vector Autoregression (VAR)[24]**: Classical time series models that provide a baseline for forecasting.
- **Prophet[36]**: A robust time series forecasting model designed to capture seasonal patterns and special events.
- **Long Short-Term Memory networks (LSTMs[23])**: Deep learning models adept at sequential data analysis.
- **Gated Recurrent Unit (GRU[23])**: Another deep learning architecture tailored for time series forecasting.

- **DeepAREstimator[25]**: Our proprietary deep learning model, engineered to tackle High Cardinality Multivariate Multi Time Series datasets.

These models are thoughtfully selected to address the unique challenges of daily COVID-19 trends and time series dynamics.

6.2.3 Training Environment

Rigorous experiments demand a controlled and consistent environment to ensure the credibility and reliability of our findings. Our training environment is as follows:

- **Operating System**: A 64-bit Debian GNU/Linux 9.11 operating system was chosen as the foundation for our experiments.
- **Hardware**: The experiments were executed on a computing infrastructure featuring an Intel (R) Xeon (R) Gold CPU @ 2.20GHz and 16 GB of RAM.
- **GPU Acceleration**: For deep learning model training, we harnessed the power of an NVIDIA Tesla K80 GPU, equipped with 12GB of memory.
- **Deep Learning Framework**: Our deep learning models were implemented and trained using the PyTorch framework.

This controlled environment is pivotal in ensuring the reproducibility and reliability of our experiments.

6.2.4 Output

The output of our experiments consists of vital performance metrics that gauge the efficacy of our Time Series Framework. The following output metrics were assessed:

- **Daily Cases**
- **Daily Deaths**
- **Daily Hospitalizations**

These metrics serve as the yardstick against which we measure the predictive capabilities of our models and the robustness of our Time Series Framework.

In the subsequent sections, we delve into the results, findings, and insights derived from these meticulously designed experiments, shedding light on the effectiveness of our Time Series Framework in the realm of COVID-19 epidemic trend prediction.

6.3 Evaluation Metrics

In assessing the performance of our predictive models within the context of our framework, it is essential to employ a set of rigorous evaluation metrics that provide insights into the quality and accuracy of predictions. Here, we discuss and elaborate on the primary metrics we have employed to gauge the effectiveness of our models.

6.3.1 R-squared (R^2)

R-squared, also known as the coefficient of determination, serves as a fundamental metric for assessing the quality of regression models. It quantifies the proportion of the variance in the dependent variable (the variable being predicted) that can be explained by the independent variables (the features utilized for prediction). The R-squared value typically ranges from 0 to 1, with the following interpretations:

$$R^2 = 1 - \frac{\sum(Predicted - Actual)^2}{\sum(Actual - \overline{Actual})^2}$$

Here, "Predicted" represents the predicted value, "Actual" corresponds to the true (observed) value, and $\overline{Actual}$ is the mean of the actual values. The R-squared value typically ranges from 0 to 1, with interpretations as mentioned earlier. A higher R-squared value indicates a better fit of the model to the data and its ability to explain a significant portion of the variation in the target variable.

- $R^2 = 0$: The model fails to explain any variance in the dependent variable, indicating a poor fit.

- $R^2 = 1$: The model perfectly explains all the variance in the dependent variable, representing a perfect fit.

A higher R-squared value signifies that the model effectively captures a significant portion of the variation in the target variable. Therefore, when the objective is to elucidate the variance in the target variable, a higher R^2 value is sought. However, it's crucial to note that in some instances, particularly when the model's fit is worse than a basic horizontal line (the "null model"), the R-squared value can assume negative values.

6.3.2 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) serves as a pivotal metric to assess the average difference between the predicted values and the actual values. It is calculated by taking the square root of the mean of the squared differences between the predictions and the true values, as defined by the formula:

$$RMSE = \sqrt{\frac{\sum(Predicted - Actual)^2}{N}}$$

Here, "Predicted" represents the predicted value, "Actual" corresponds to the true (observed) value, and "N" is the number of data points.

RMSE is particularly sensitive to outliers and places a higher penalty on larger prediction errors. A lower RMSE value indicates better model performance as it signifies that the model's predictions closely align with the actual values. Consequently, when the primary goal is to achieve accurate predictions, a lower RMSE value is preferred.

6.3.3 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is akin to RMSE but differs in its approach to error calculation. Instead of squaring the errors, it takes the absolute value of the differences between the predicted values and the actual values. MAE is computed using the formula:

$$MAE = \frac{\sum |Predicted - Actual|}{N}$$

In this formula, "Predicted" represents the predicted value, "Actual" corresponds to the true (observed) value, and "N" is the number of data points.

MAE is relatively less sensitive to outliers compared to RMSE since it does not involve squaring the errors. Like RMSE, lower MAE values indicate superior model performance, signifying that the model's predictions exhibit smaller absolute differences from the actual values. Therefore, when the primary objective is to achieve accurate predictions, a lower MAE value is sought.

6.3.4 Max Absolute Error (MaxAE)

The Max Absolute Error (MaxAE) is defined as the maximum absolute difference between the predicted values and the actual values across the dataset. It is computed using the following formula:

$$MaxAE = \max |Predicted - Actual|$$

In this formula, "Predicted" represents the predicted value, and "Actual" corresponds to the true (observed) value. The MaxAE metric provides insights into the largest prediction error encountered in the model's performance. A lower MaxAE value signifies better model accuracy, implying that even the worst-case errors are relatively small.

6.4 Results

6.4.1 Critical Feature Identifications

The most critical aspects of our analysis—the identification of key features that play a pivotal role in predicting the daily cases, daily hospitalizations, and daily deaths. These critical features, carefully selected through a rigorous evaluation process, serve

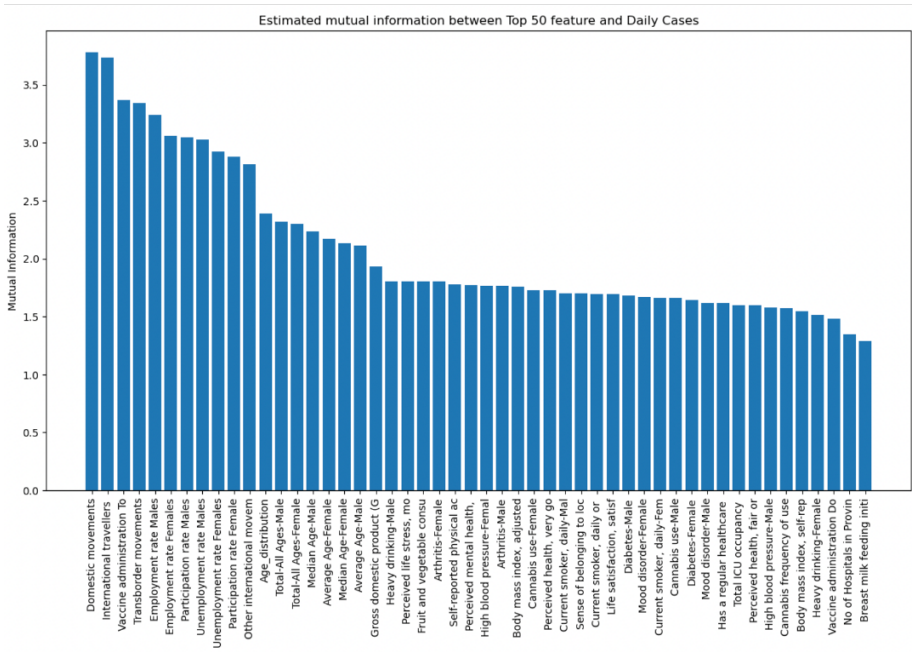


Fig. 6.4.1: Critical features for Daily Cases

as the cornerstone of our predictive models. Understanding their significance not only sheds light on the factors influencing the course of the COVID-19 pandemic but also holds the potential to inform targeted interventions and strategies for managing its impact. As we navigate through the results, we invite you to explore the intricate web of relationships between these features and the epidemic trends, providing a comprehensive view of the data-driven insights that underpin our framework's predictive capabilities.

6.4.2 Regression Model Results

In our extensive evaluation of regression models within the context of our regression framework, we have witnessed a remarkable performance that sets our framework apart from other regression models, including K-Nearest Neighbors (KNN) and Random Forest, across various critical aspects of model assessment. These aspects encompass essential metrics such as R-squared (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Max Absolute Error (MaxAE). Our regression framework excelled in all of these metrics, signifying its superior predictive capabili-

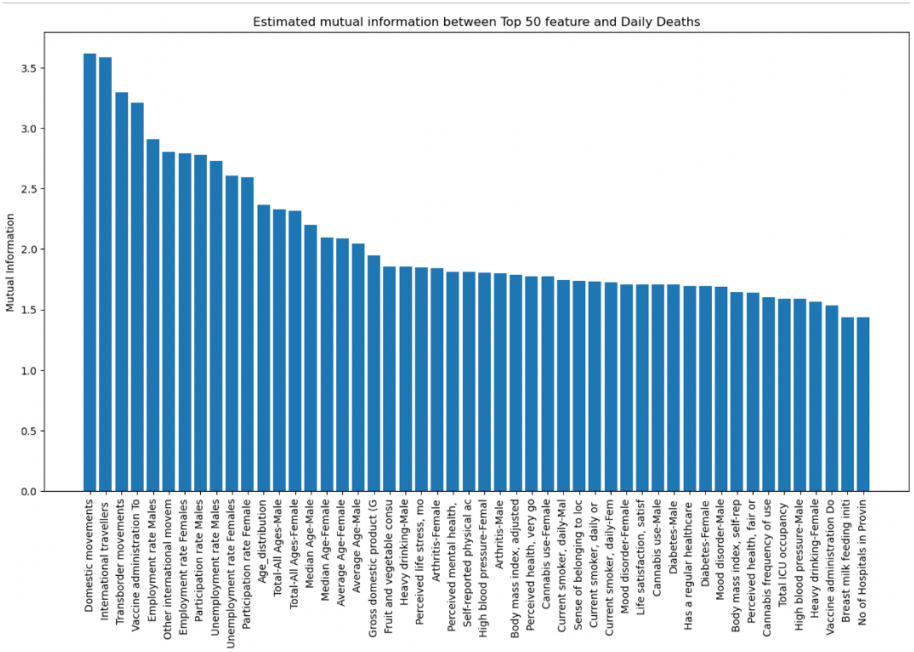


Fig. 6.4.2: Critical features for Daily Deaths

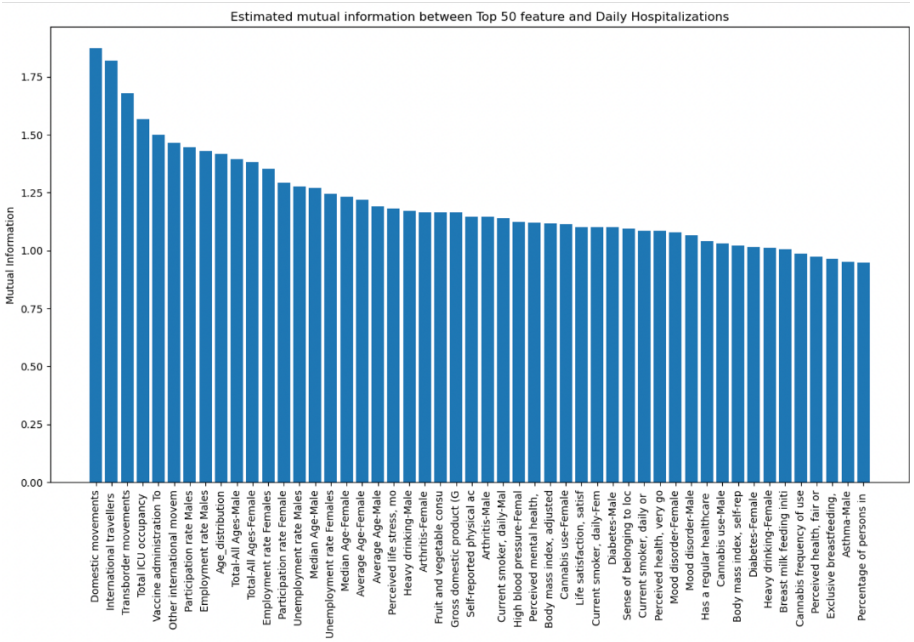


Fig. 6.4.3: Critical features for Daily Hospitalisations

ties when compared to the alternative models.

Specifically, our framework achieved a higher R^2, indicating that it comprehensively explained a significant portion of the variance in the dependent variable. It also exhibited lower values for RMSE and MAE, signifying that its predictions were consistently closer to the actual values, an essential characteristic when aiming for precise and accurate predictions. Moreover, the MaxAE values were kept in check, further underlining the framework's ability to minimize extreme prediction errors.

After a series of meticulous experiments, we found that the Neural Network model with four hidden layers emerged as the top performer. This configuration, when combined with a batch size of 8, the Adam optimizer, a dropout rate of 0.5, and a training duration spanning 150-250 epochs, consistently delivered the most robust results. The choice of the Rectified Linear Unit (ReLU) as the activation function also played a crucial role in optimizing our framework's predictive accuracy, particularly concerning the total number of cases, total hospitalizations, and total deaths. This combination of architectural elements represents a powerful recipe for successful predictions in the context of our regression framework.

While our regression framework has proven its mettle in various aspects, it's worth noting that it does demand relatively more time in terms of computational resources due to its deep neural network architecture. However, this investment in time is well-justified, given the significant enhancements in predictive accuracy and the crucial insights it provides in understanding and tackling the challenges posed by the COVID-19 pandemic.

After conducting each experiment five times, we calculated the mean values of the evaluated metrics to obtain a more robust and reliable measure of the model's performance. This approach not only mitigates the impact of outliers but also ensures that the reported results are a representative measure of the model's capabilities. By considering the mean values, we aim to provide a comprehensive and stable assessment of the model's predictive accuracy and generalizability across different scenarios.

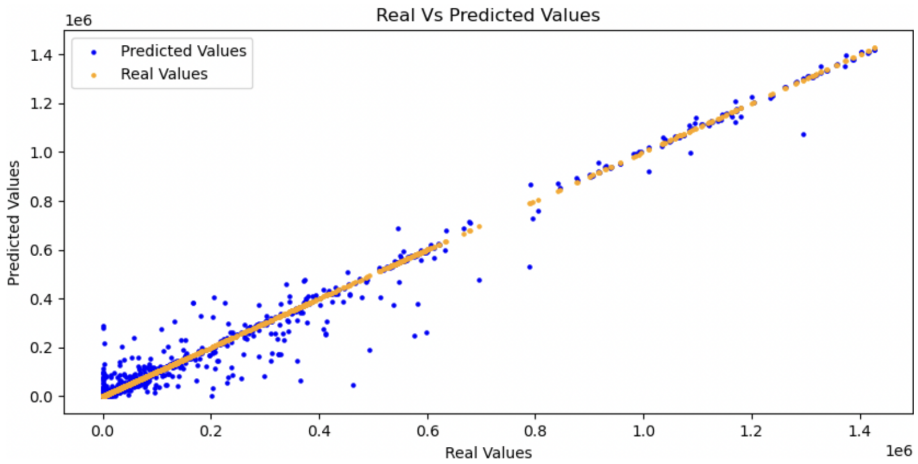


Fig. 6.4.4: Prediction Of Total Daily Cases KNN

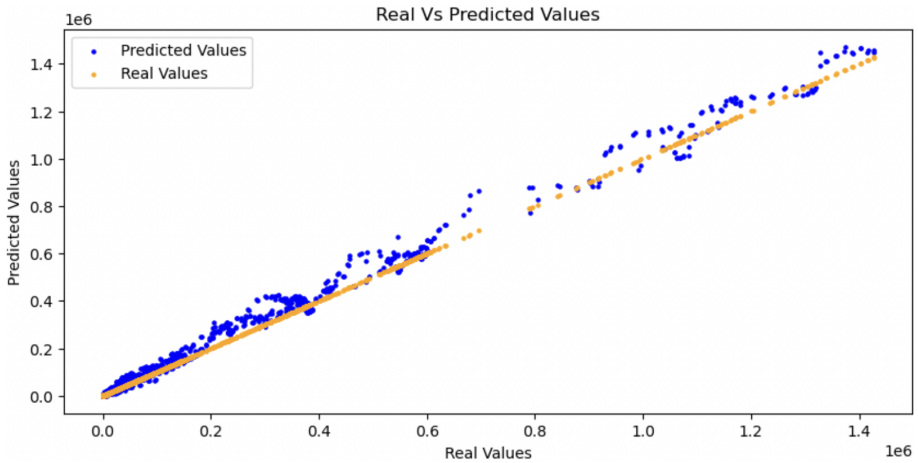


Fig. 6.4.5: Prediction Of Total Daily Cases Random Forest

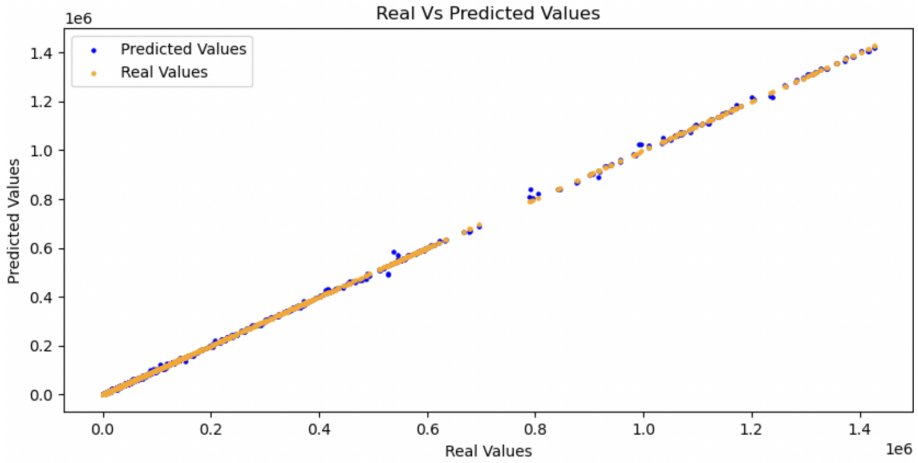


Fig. 6.4.6: Prediction Of Total Daily Cases Regression Framework

Regression Algorithms	**R^2**	**RMSE**	**MAE**	**MaxAE**	**Training Time**
KNN	0.79	32168.6486	9352.1366	417673.0000	**0.04**
Random Forest	0.87	16418.0310	8078.9526	133431.4375	36.36
Regression Framework	**0.99**	**2930.4405**	**924.9775**	**49195.6999**	1515.69

Table 6.4.1: Predicting Total Daily Cases

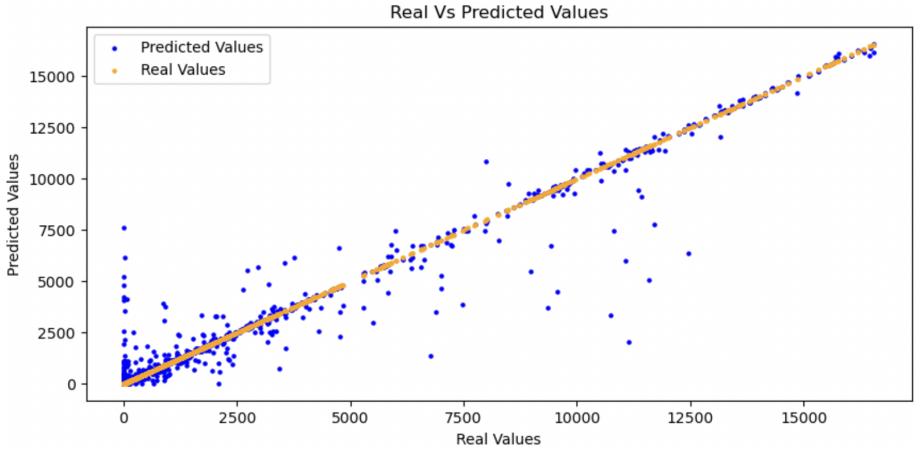


Fig. 6.4.7: Prediction Of Total Daily Deaths KNN

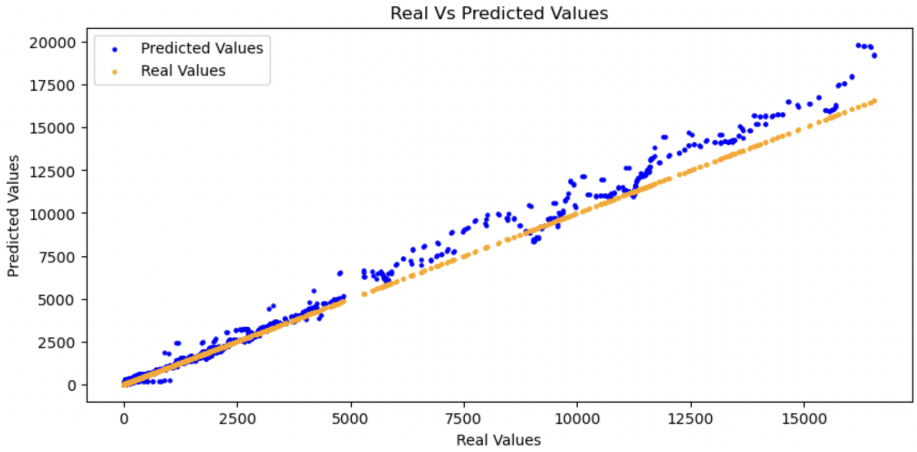


Fig. 6.4.8: Prediction Of Total Daily Deaths Random Forest

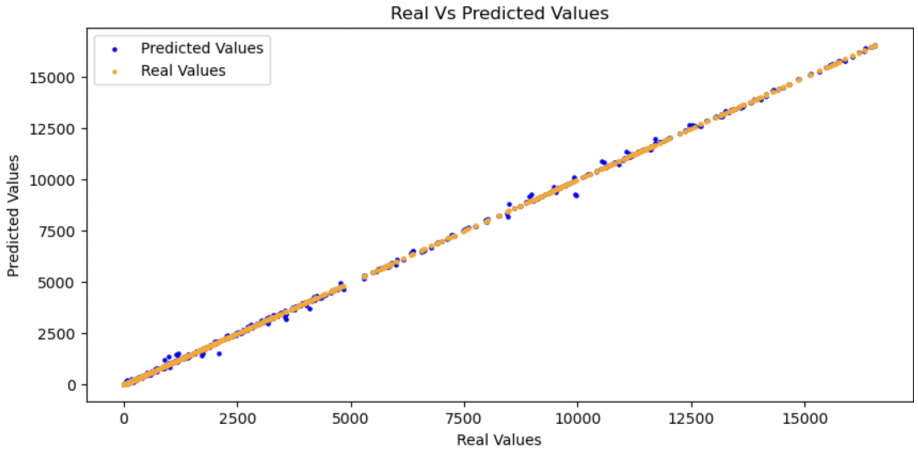


Fig. 6.4.9: Prediction Of Total Daily Deaths Regression Framework

Regression Algorithms	**R^2**	**RMSE**	**MAE**	**MaxAE**	**Training Time**
KNN	0.81	590.0135	135.9663	9078.0000	**0.04**
Random Forest	0.88	198.5118	91.7215	1758.4102	26.47
Regression Framework	**0.99**	**41.9891**	**11.8251**	**747.0900**	1455.26

Table 6.4.2: Predicting Total Daily Deaths

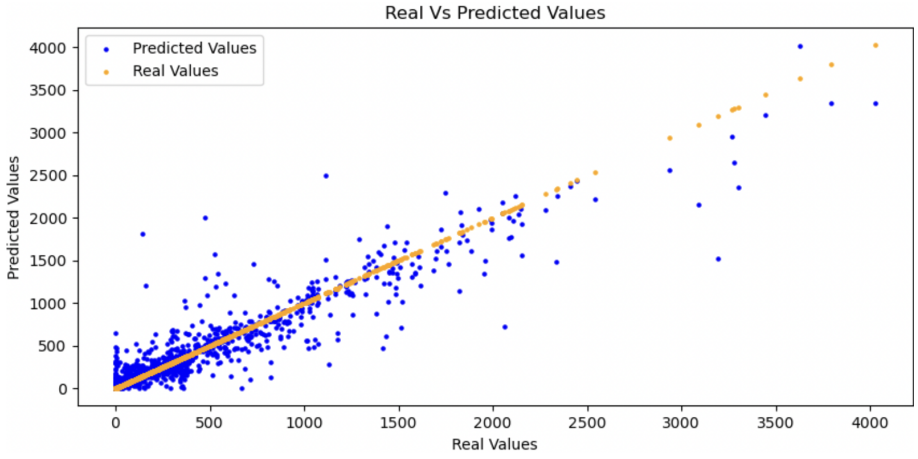


Fig. 6.4.10: Prediction Of Total Daily Hospitalisations KNN

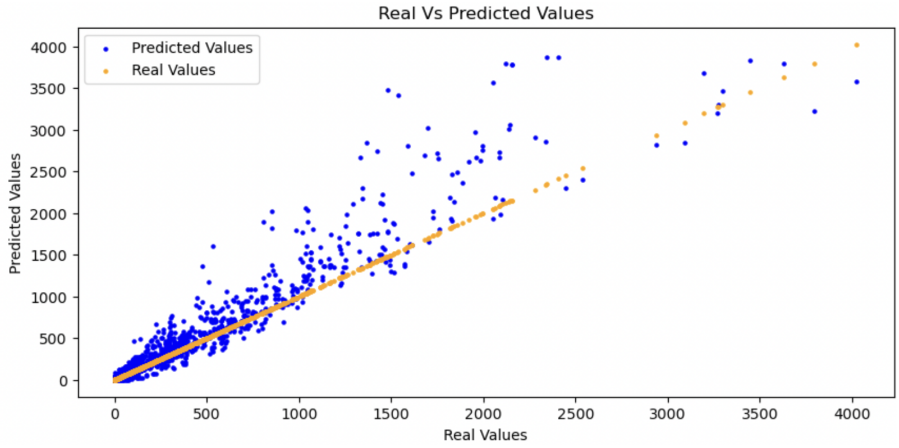


Fig. 6.4.11: Prediction Of Total Daily Hospitalisations Random Forest

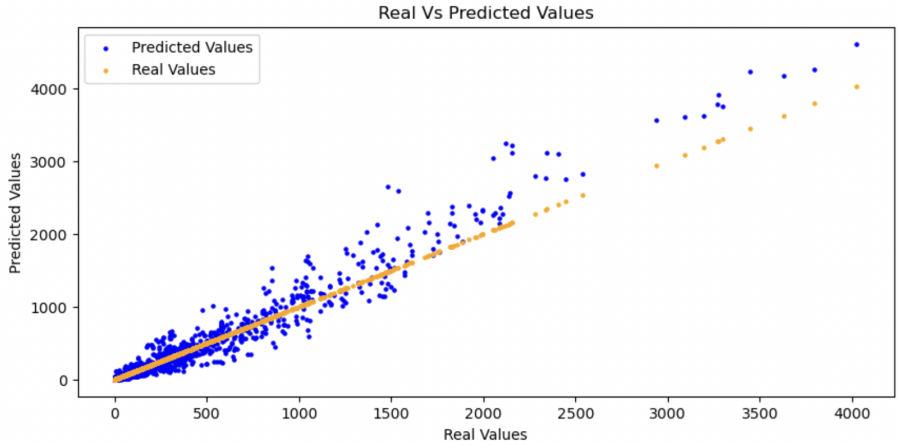


Fig. 6.4.12: Prediction Of Total Daily Hospitalisations Regression Framework

Regression Algorithms	**R^2**	**RMSE**	**MAE**	**MaxAE**	**Training Time**
KNN	0.78	590.0135	135.9663	9078.0000	**0.04**
Random Forest	0.85	68.4125	31.0935	1131.8500	32.63
Regression Framework	**0.98**	**61.1323**	**16.0368**	**499.9365**	1539.64

Table 6.4.3: Predicting Total Daily Hospitalisations

6.4.3 Time Series Model Results

When assessing the performance of time series models within our comprehensive time series framework, we conducted a thorough evaluation to gauge the effectiveness of various approaches. One of our primary objectives was to compare the results of traditional ARIMA and VAR-based models with those obtained from a specialized framework that harnessed the power of deep neural networks to generate essential encodings. These encodings were subsequently employed as inputs for alternative time series models, including Prophet, LSTMs, and the DeepAREstimator.

Our evaluation was multi-faceted, encompassing a range of metrics to comprehensively measure the models' predictive capabilities. Among the critical metrics considered were R-squared (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Maximum Absolute Error (MaxAE). These metrics provided insight into the models' accuracy, precision, and ability to capture the underlying patterns and variations in the data.

R^2, or the coefficient of determination, offered a glimpse into how well each model explained the variance in the target variables. RMSE quantified the average difference between predicted and actual values, while MAE provided a measure of the mean absolute differences. Finally, MaxAE highlighted the maximum error encountered in the predictions.

In addition to these metrics, we also examined the time factor, recognizing the importance of model efficiency in real-world applications. The time taken for model training and prediction was a crucial consideration, as it directly impacts the feasibility of deploying these models in real-time scenarios.

Through a systematic and rigorous assessment, we aimed to uncover which models excelled in capturing the intricate dynamics of the time series data associated with COVID-19 epidemic trends. This evaluation not only sheds light on the individual strengths and weaknesses of each model but also aids in identifying the most promising candidates for accurate and efficient predictions.

In the forthcoming results, we present a comprehensive analysis of our findings,

showcasing how each model performed across these key evaluation metrics and highlighting their respective advantages and limitations. This exploration is instrumental in providing a clear roadmap for selecting the most suitable time series model within our framework for the prediction of COVID-19 epidemic trends.

6.4.3.1 Prophet Model Configuration for Time Series Framework

In our pursuit of finding the most suitable model for our time series framework, Prophet, a robust forecasting tool developed by Facebook, emerged as a notable contender. Prophet is celebrated for its ability to handle time series data with remarkable accuracy and simplicity. During our evaluation, one of the noteworthy findings was that the default configuration of Prophet yielded the best results for our framework, showcasing its robust performance in capturing the intricacies of COVID-19 epidemic trends.

Prophet, by design, is equipped to handle time series data with a particular emphasis on capturing seasonality, holidays, and other recurrent patterns. Its default configuration settings have been meticulously crafted to provide an excellent starting point for time series analysis. These defaults encompass various aspects, including the treatment of seasonality, holiday effects, and trend flexibility.

One of the strengths of Prophet lies in its automatic detection of seasonality, making it an ideal choice for time series data that exhibit regular patterns over time. The default settings of Prophet are tailored to identify such seasonality components and adjust the model accordingly, ensuring that these patterns are well-captured during the prediction process.

Another noteworthy feature of Prophet is its consideration of holiday effects. It can incorporate holidays or special events as additional input features, enabling the model to account for sudden, non-seasonal changes in the time series. This inclusion of holidays is pivotal, especially in scenarios like the prediction of COVID-19 epidemic trends, where policy changes and public holidays can significantly impact the data.

Furthermore, Prophet's default configuration includes provisions for adjusting the flexibility of the underlying trend. The default settings automatically determine the

best balance between seasonality, holidays, and the overall trend, allowing the model to effectively capture the temporal dynamics of the data.

Our rigorous evaluation of Prophet revealed that these default settings consistently outperformed alternative configurations. The default configuration proved to be not only highly accurate but also robust, requiring minimal parameter tuning. This simplicity and effectiveness make it a valuable asset within our time series framework.

The following section presents the comprehensive results of our experiments with Prophet as part of our time series framework. The default configuration's performance is showcased across a range of evaluation metrics, highlighting its superiority in capturing the nuanced dynamics of COVID-19 epidemic trends. Through our detailed analysis and thorough experimentation, the default settings of Prophet have demonstrated their prowess, reaffirming their status as an ideal choice for our predictive endeavors.

6.4.3.2 Deep Learning Models(LSTMs/GRU) Configuration for Time Series Framework

In our pursuit of achieving the most accurate and effective model for time series analysis within the COVID-19 epidemic trend prediction framework, we conducted a meticulous examination of various architectural configurations. Our focus was particularly on Long Short-Term Memory (LSTM) networks, renowned for their exceptional ability to capture sequential dependencies in time series data.

One of the significant milestones in our experimentation was the development of a deep LSTM architecture with four hidden layers. Each of these layers played a pivotal role in comprehending the intricate temporal patterns that characterize COVID-19 epidemic trends. We adopted the Adam optimizer to facilitate the convergence of our model, working alongside a window length of 10 for features. In addition, the use of past-day predictions as input features for current-day predictions proved to be a crucial component of our framework.

Further fine-tuning of our LSTM network involved the careful selection of hyperparameters. Our choice of a batch size of 4, coupled with a relatively high number

of epochs (1000), allowed our model to iteratively learn and adapt to the temporal complexities within the data. To prevent overfitting and ensure robust generalization, a dropout rate of 0.3 was employed, striking a balance between learning from data and preventing model complexity.

Activation functions are central to the success of deep learning networks, and we found that Leaky Rectified Linear Unit (LeakyReLU) activation functions worked exceptionally well in tandem with the LSTM architecture. These activation functions allowed for the exploration of both linear and non-linear relationships within the time series data, further enhancing the model's capacity to capture and predict COVID-19 epidemic trends.

One noteworthy observation from our experimentation was the superior performance of LSTMs compared to Gated Recurrent Units (GRUs). While both LSTM and GRU networks are proficient at modeling sequential data, the LSTM architecture demonstrated greater effectiveness in this context. This observation is instrumental in guiding our model selection process, ensuring that we employ the most suitable deep learning architecture within our time series framework.

The combination of the above-mentioned architectural elements and hyperparameter settings led to remarkable results in predicting COVID-19 epidemic trends. As we delve into the results section, we will present a comprehensive analysis, providing insight into the models' performance and highlighting the potential of LSTM networks within our framework for accurate and reliable time series prediction.

6.4.3.3 DeepAREstimator Model Configuration for Time Series Framework

In the realm of time series modeling for the COVID-19 epidemic trend prediction framework, we explored a wide array of powerful models, including the DeepAR estimator. DeepAR is a neural network-based approach designed for probabilistic time series forecasting. Through meticulous experimentation, we identified the optimal configuration that enabled the DeepAR model to achieve exceptional predictive accuracy.

One key aspect of our DeepAR experimentation involved setting the context length to 10 for features. This context length, in essence, determines how many past observations the model considers when making predictions for the current day. In our comprehensive study, a context length of 10 emerged as the most effective choice, allowing the model to capture the essential temporal dependencies and patterns within the COVID-19 epidemic trends.

Another crucial parameter we fine-tuned was the number of layers within the DeepAR architecture. Our results demonstrated that a three-layered model configuration offered the best performance. This choice of the number of layers played a pivotal role in balancing model complexity while preserving its capacity to understand intricate temporal relationships in the data.

We further optimized our DeepAR model by conducting experiments with different hyperparameter settings. Notably, we found that running the model for 400 epochs yielded the most accurate predictions. An epoch represents one complete cycle through the entire training dataset. This extensive training duration enabled the model to converge to a state of high accuracy and predictive power, making it an ideal choice for COVID-19 epidemic trend forecasting.

It's worth noting that, in our experimentation, we adhered to the default settings for various other parameters that DeepAR offers, such as batch size, the choice of optimizer, and dropout rates. These default settings have been meticulously crafted by the developers of the DeepAR framework and have proven to work effectively in the context of our COVID-19 epidemic trend prediction.

The outcome of our comprehensive experiments with DeepAR is a highly accurate, powerful model that provides probabilistic forecasts for COVID-19 epidemic trends. As we delve into the results section, we will present in-depth insights into the performance of the DeepAR estimator and demonstrate how it contributes to our time series framework's capacity to predict COVID-19 epidemic trends with precision and reliability.

Now, with Alberta Province and Daily Cases Prediction problem we will be looking at a complete example to understand our timeline and prediction views for better

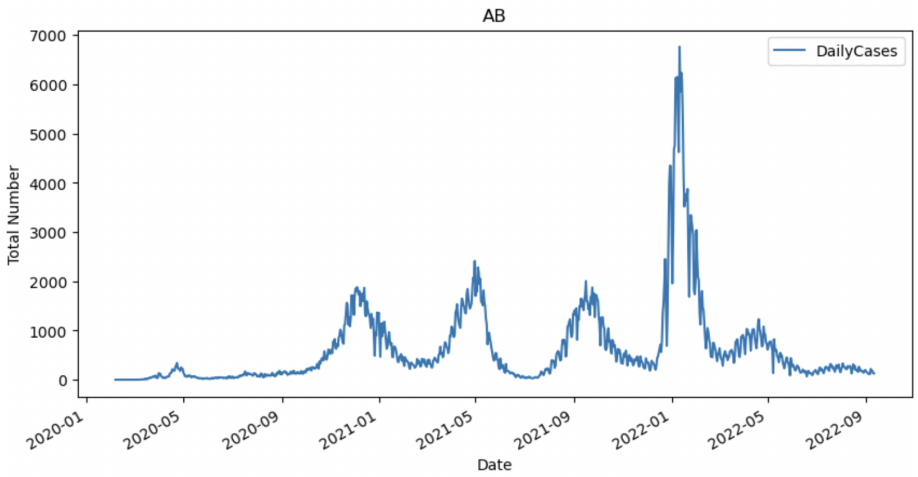


Fig. 6.4.13: Daily Cases Full Time Line (Region Alberta)

understanding of Results

6.4.4 Time Series Model Results Daily Cases

Results for Daily Cases are discussed below

In Fig 6.4.13 We have full time line view of all the points and corresponding Daily Cases for Alberta Province

In Fig 6.4.14 We have Training Data time line view of all the points and corresponding Daily Cases for Alberta Province. This is the data that has been used to train our different time series models such as Prophet, LSTMs(Deep Learning), DeepAREstimator after encodings have been received for each date from Feed forward neural network of our regression framework.

In Fig 6.4.15 We have Validation Data time line view of all the points and corresponding Daily Cases for Alberta Province. This is the data that has been used to validate our different time series models such as Prophet, LSTMs(Deep Learning), DeepAREstimator after encodings have been received for each date from Feed forward neural network of our regression framework.

In Fig 6.4.16 We have Test Data time line view of all the points and corresponding Daily Cases for Alberta Province. This is the data that has been used to test our different time series models such as Prophet, LSTMs(Deep Learning), DeepAREs-

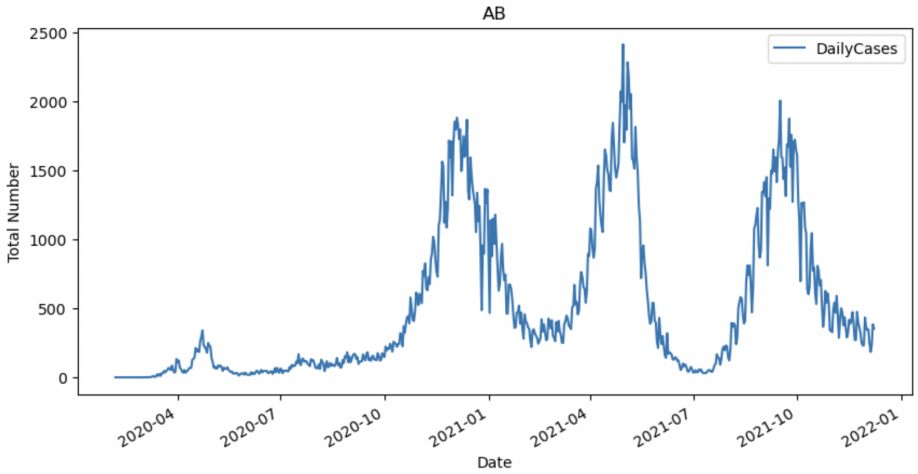


Fig. 6.4.14: Daily Cases Train Time Line (Region Alberta)

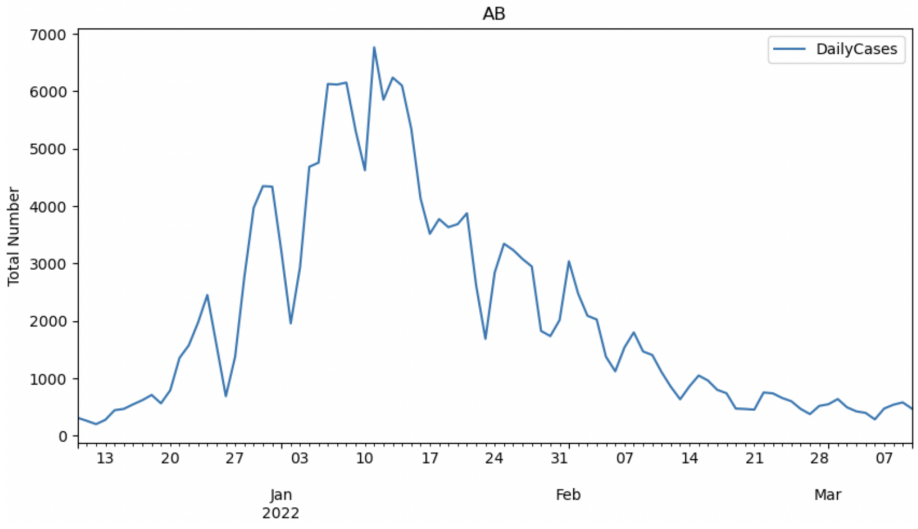


Fig. 6.4.15: Daily Cases Validation Time Line (Region Alberta)

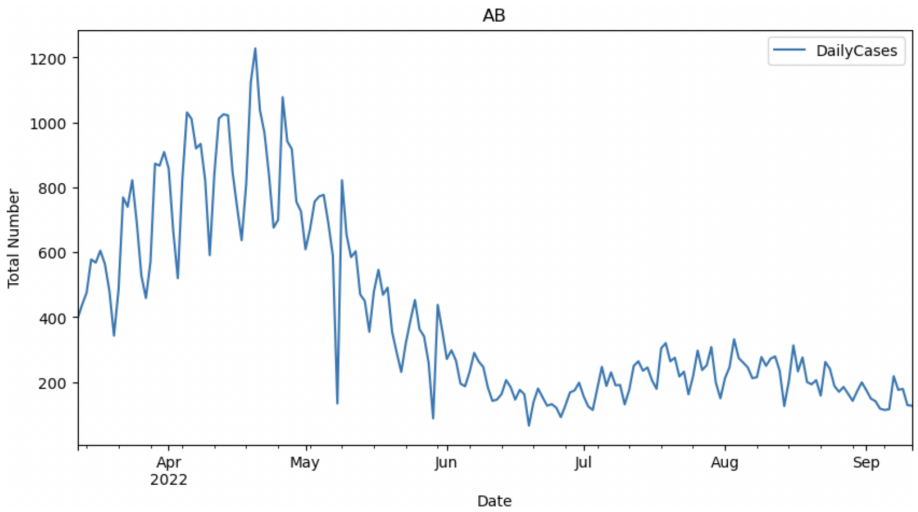


Fig. 6.4.16: Daily Cases Test Time Line (Region Alberta)

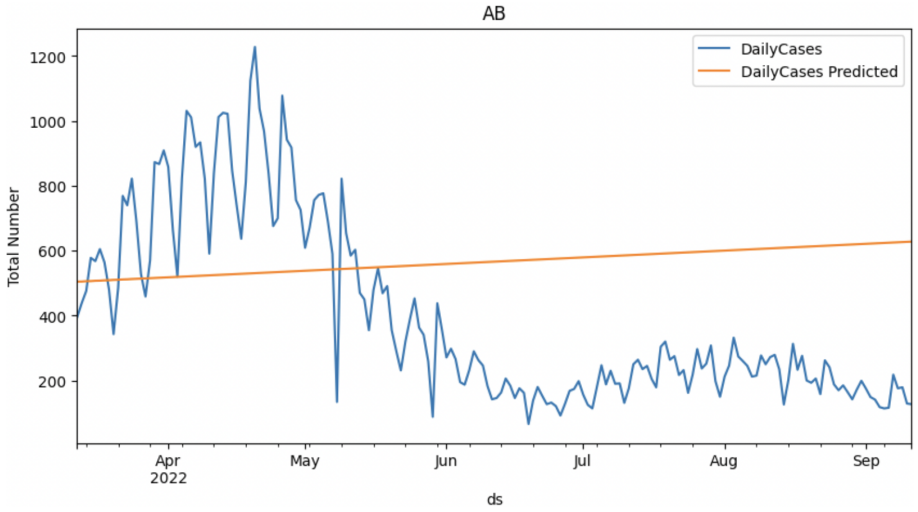


Fig. 6.4.17: Prediction Of Daily Cases Arima Model (Region Alberta)

timator after encodings have been received for each date from Feed forward neural network of our regression framework.

In Fig 6.4.17 We have Test Data line Prediction view of all the points and corresponding Daily Cases for Alberta Province. This is prediction done by ARIMA model after it was trained and validated on respective data points.

In Fig 6.4.18 We have Test Data line Prediction view of all the points and corresponding Daily Cases for Alberta Province. This is best prediction done by our Framework which in this case used LSTMs based Deep Learning model after the same was trained and validated on respective data points.

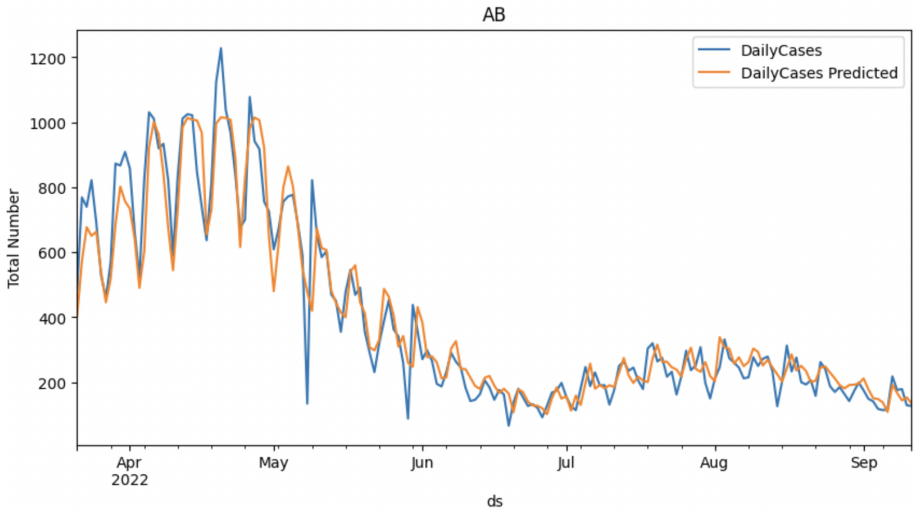


Fig. 6.4.18: Prediction Of Daily Cases Deep Learning Time Series Framework-LSTM Model (Region Alberta)

The Alberta Province exhibited optimal performance on the testing data when predicting daily cases. The training and testing datasets closely followed similar distributions, facilitating the model's ability to generalize effectively.

In Fig 6.4.19 We have Test Data line Prediction view of all the points and corresponding Daily Cases for Alberta Province. This is prediction done by our Framework using DeepAREstimator model after the same was trained and validated on respective data points.

In Fig 6.4.20 We have Full Time line Prediction view of all the points and corresponding Daily Cases for Alberta Province. This is prediction done by our Framework after the same was trained and validated on respective data points.

In Fig 6.4.21 We have Full Time line Prediction view with 95% boundary of all the points and corresponding Daily Cases for Alberta Province. This is prediction done by our Framework after the same was trained and validated on respective data points.

In Fig 6.4.22 We have full time line view of all the points and corresponding Daily Cases for Quebec Province

In Fig 6.4.23 We have Full Time line Prediction view of all the points and corresponding Daily Cases for Quebec Province. This is prediction done by our

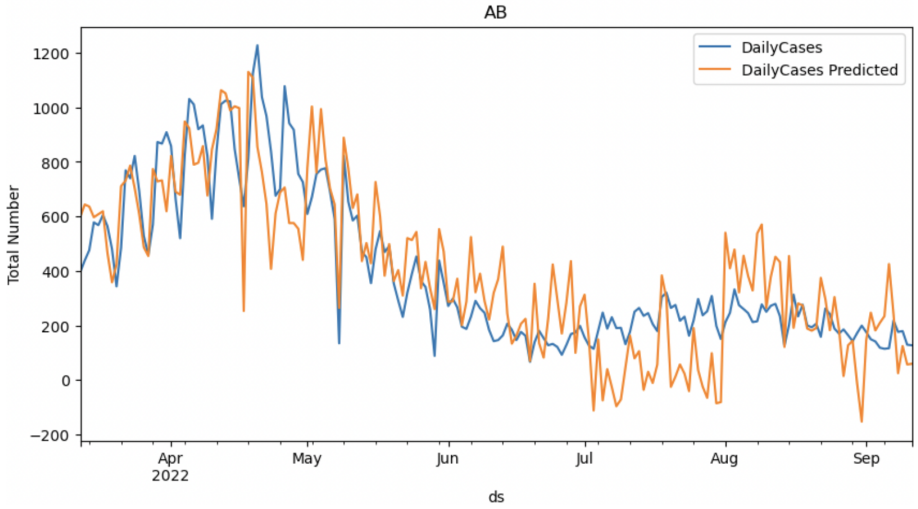


Fig. 6.4.19: Prediction Of Daily Cases Deep Learning Time Series Framework-DeepAREstimator (Region Alberta)

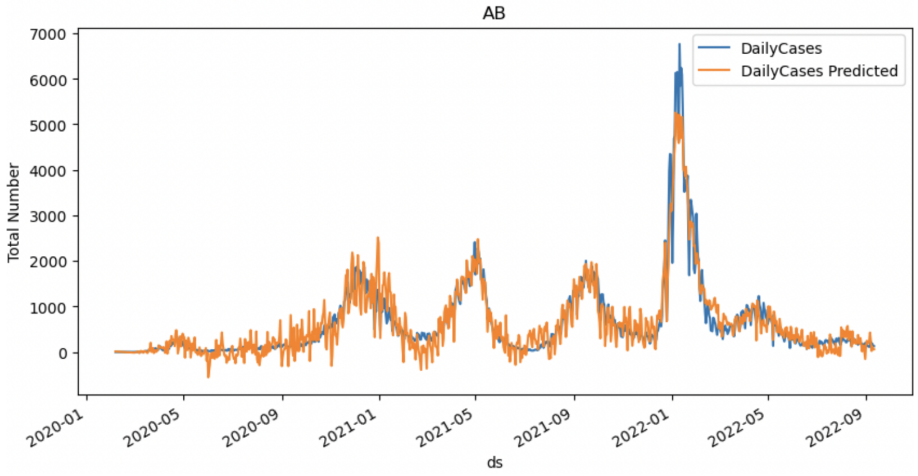


Fig. 6.4.20: Prediction Of Daily Cases Deep Learning Time Series Framework Full Time Line (Region Alberta)

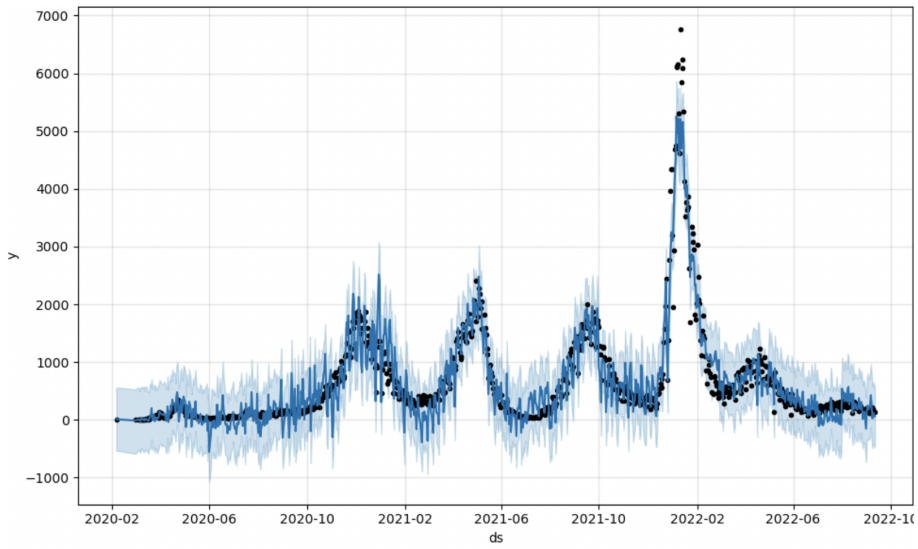


Fig. 6.4.21: Prediction Of Daily Cases Deep Learning Time Series Framework-DeepAREstimator Full Time Line Percentile View(Region Alberta)

Time Series Models	**R^2**	**RMSE**	**MAE**	**MaxAE**	**[1539.64] + Training Time**
ARIMA	-0.55	350.0980	320.0554	697.1910	**1.25**
DL-TS Framework-Prophet	0.90	171.4782	135.5246	430.6880	2.53
DL-TS Framework-LSTMs	**0.95**	**79.3224**	**53.1890**	**401.9676**	1410.65
DL-TS Framework-DeerAREstimator	0.92	88.3984	58.2643	418.4246	338.46

Table 6.4.4: Predicting Daily Cases (Region Alberta)

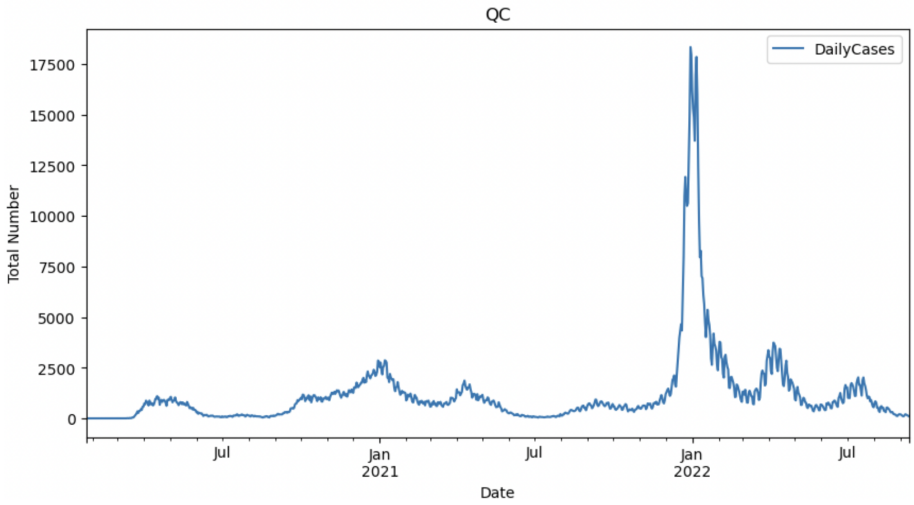


Fig. 6.4.22: Daily Cases Full Time Line (Region Quebec)

Time Series Models	**R^2**	**RMSE**	**MAE**	**MaxAE**	**[1539.64] + Training Time**
ARIMA	-0.13	889.1554	717.3580	2604.8109	**1.25**
DL-TS Framework-Prophet	0.88	505.8036	377.3975	1577.3555	2.53
DL-TS Framework-LSTMs	**0.92**	**305.8036**	**223.1890**	**1213.9676**	1415.65
DL-TS Framework-DeerAREstimator	0.90	328.3984	246.4674	1346.4246	338.46

Table 6.4.5: Predicting Daily Cases (Region Quebec)

Framework after the same was trained and validated on respective data points.

In Fig 6.4.24 We have full time line view of all the points and corresponding Daily Cases for Ontario Province

In Fig 6.4.25 We have Full Time line Prediction view of all the points and corresponding Daily Cases for Ontario Province. This is prediction done by our Framework after the same was trained and validated on respective data points.

In Fig 6.4.26 We have Full Time line Prediction view of all the points and corresponding Daily Cases for British Columbia Province. This is prediction done by

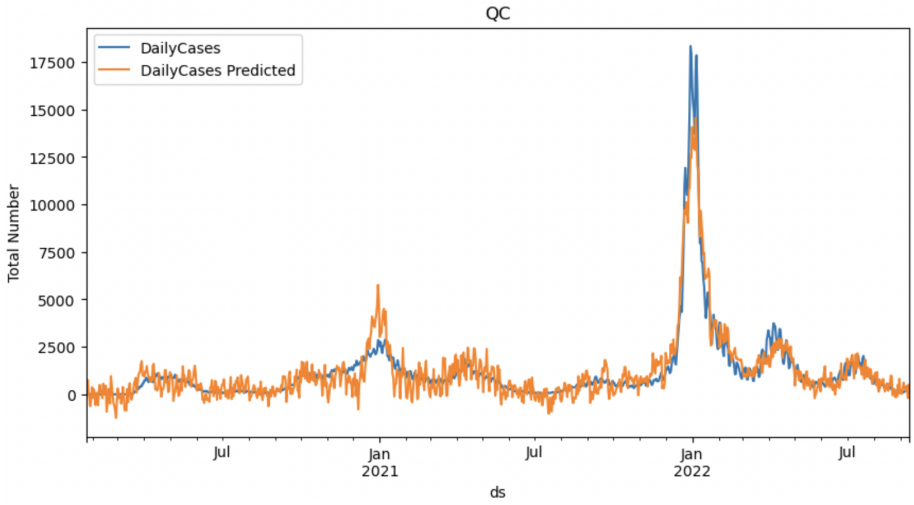


Fig. 6.4.23: Prediction Of Daily Cases Deep Learning Time Series Framework Full Time Line (Region Quebec)

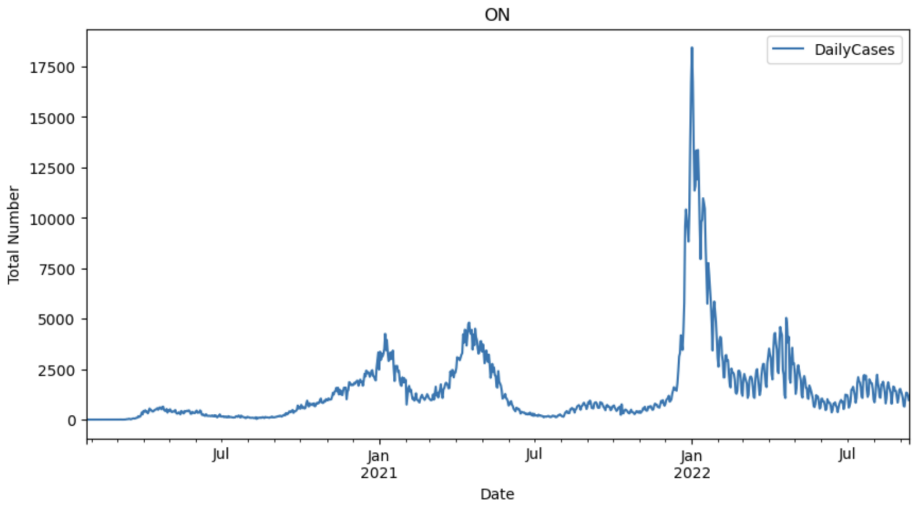


Fig. 6.4.24: Daily Cases Full Time Line (Region Ontario)

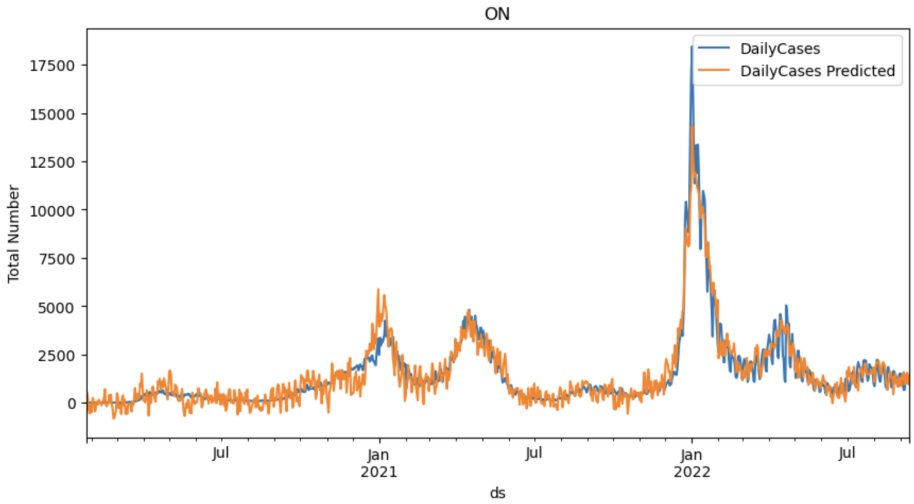


Fig. 6.4.25: Prediction Of Daily Cases Deep Learning Time Series Framework Full Time Line (Region Ontario)

Time Series Models	**R^2**	**RMSE**	**MAE**	**MaxAE**	**[1539.64] + Training Time**
ARIMA	-1.51	1477.0129	1334.4758	2450.3300	**1.25**
DL-TS Framework-Prophet	0.90	616.2956	468.2510	2391.4608	2.53
DL-TS Framework-LSTMs	**0.92**	**556.2956**	**438.4735**	**2254.1346**	1404.32
DL-TS Framework-DeerAREstimator	0.89	576.3476	442.3165	2425.4342	338.46

Table 6.4.6: Predicting Daily Cases (Region Ontario)

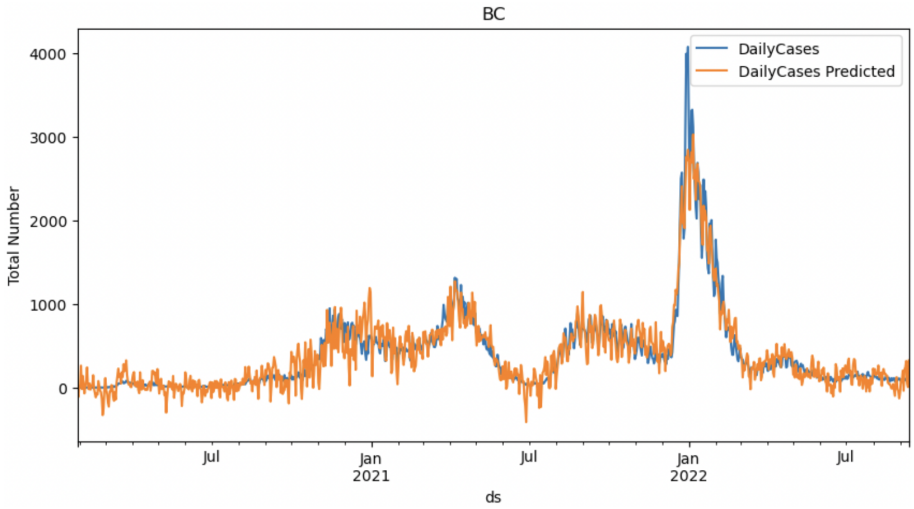


Fig. 6.4.26: Prediction Of Daily Cases Deep Learning Time Series Framework Full Time Line (Region British Columbia)

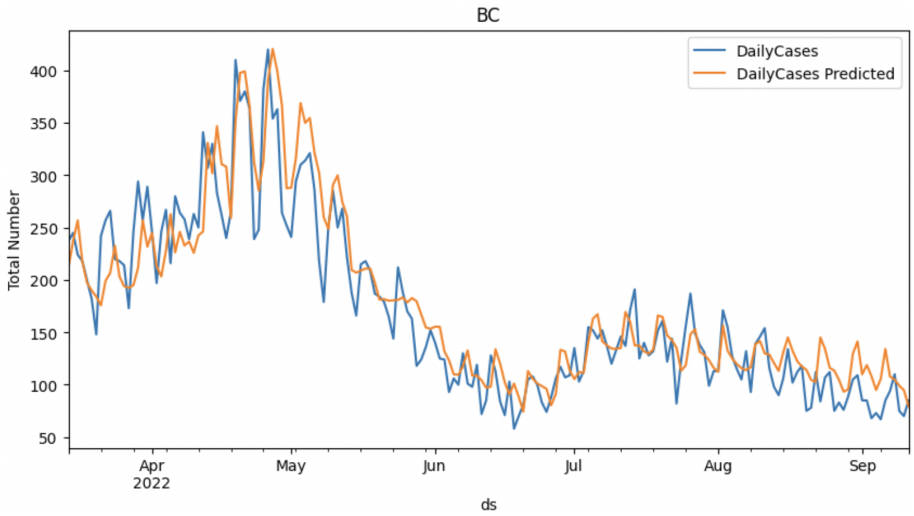


Fig. 6.4.27: Prediction Of Daily Cases Deep Learning Time Series Framework Test Time Line (Region British Columbia)

Time Series Models	**R^2**	**RMSE**	**MAE**	**MaxAE**	**[1539.64] + Training Time**
ARIMA	-5.88	217.7841	195.2429	337.9356	**1.25**
DL-TS Framework-Prophet	0.89	161.9411	113.6925	255.3342	2.53
DL-TS Framework-LSTMs	**0.96**	**31.9732**	**24.2660**	**102.4750**	1462.54
DL-TS Framework-DeerAREstimator	0.93	103.9015	84.2496	238.3783	338.46

Table 6.4.7: Predicting Daily Cases (Region British Columbia)

our Framework when using Prophet model after the same was trained and validated on respective data points.

In Fig 6.4.27 We have Test Time line Prediction view of all the points and corresponding Daily Cases for British Columbia Province. This is prediction done by our Framework using LSTMs after the same was trained on respective data points.

6.4.5 Time Series Model Results Daily Deaths

Results for Daily Deaths are discussed below

In Fig 6.4.28 We have Full Time line Prediction view of all the points and corresponding Daily Deaths for British Columbia Province. This is prediction done by our Framework when using Prophet model after the same was trained and validated on respective data points.

In Fig 6.4.29 We have Test Time line Prediction view of all the points and corresponding Daily Deaths for British Columbia Province. This is prediction done by our Framework using LSTMs after the same was trained on respective data points.

In Fig 6.4.30 We have Full Time line Prediction Percentile view of all the points and corresponding Daily Deaths for Quebec Province. This is prediction done by our Framework when using Prophet model after the same was trained and validated on

Time Series Models	**R^2**	**RMSE**	**MAE**	**MaxAE**	**[1455.26] + Training Time**
ARIMA	-1.39	14.8589	12.7556	57.9586	**1.62**
DL-TS Framework-Prophet	0.89	8.8575	5.9920	78.3149	2.46
DL-TS Framework-LSTMs	**0.92**	**7.3667**	**5.1863**	**49.5068**	1562.54
DL-TS Framework-DeerAREstimator	0.90	7.9015	6.3532	54.5215	348.46

Table 6.4.8: Predicting Daily Dealths (Region Ontario)

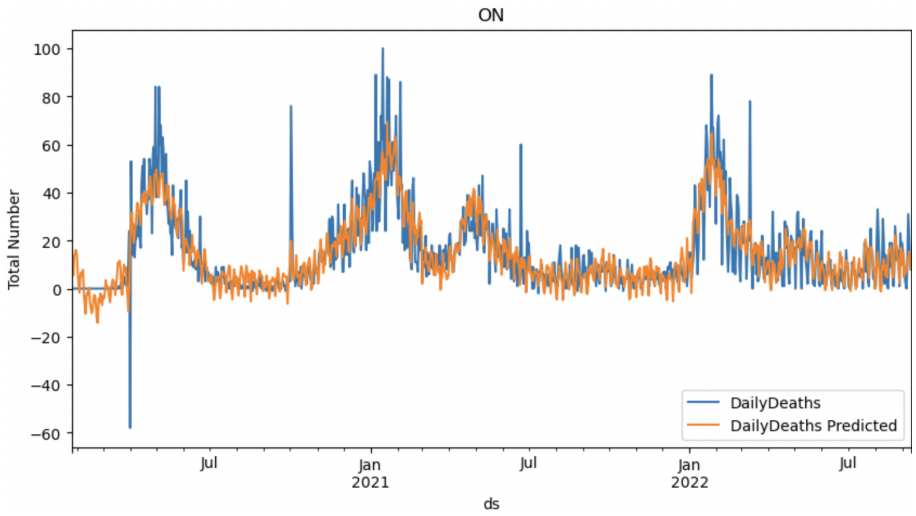


Fig. 6.4.28: Prediction Of Daily Deaths Deep Learning Time Series Framework Full Time Line (Region Ontario)

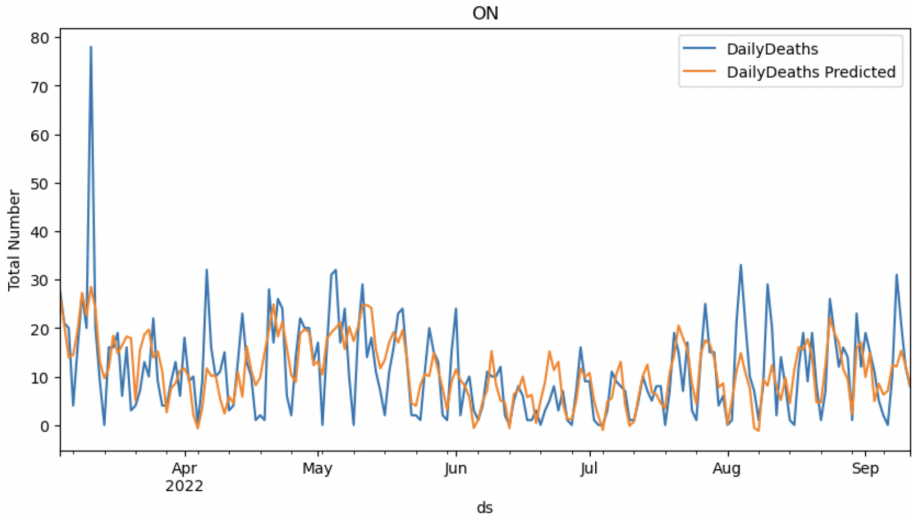


Fig. 6.4.29: Prediction Of Daily Deaths Deep Learning Time Series Framework Test Time Line (Region Ontario)

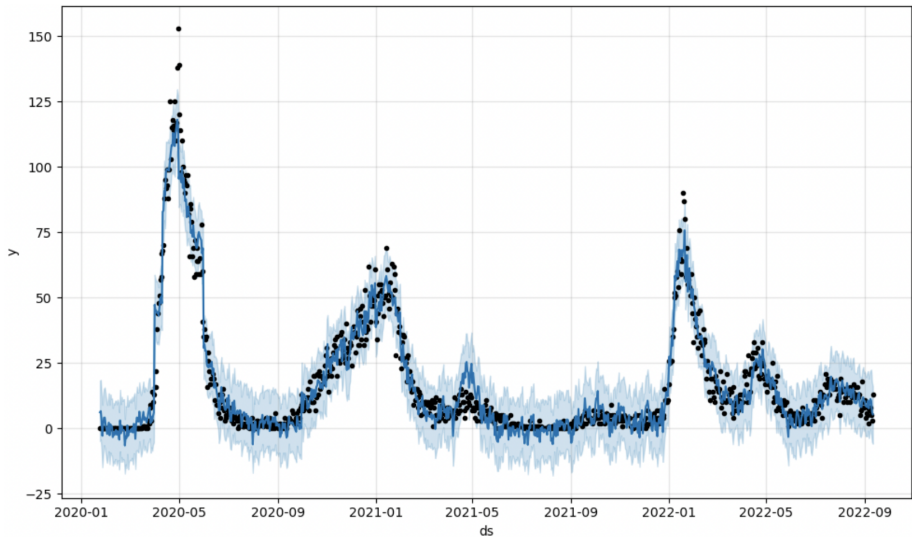


Fig. 6.4.30: Prediction Of Daily Deaths Deep Learning Time Series Framework Full Time Line (Region Quebec)

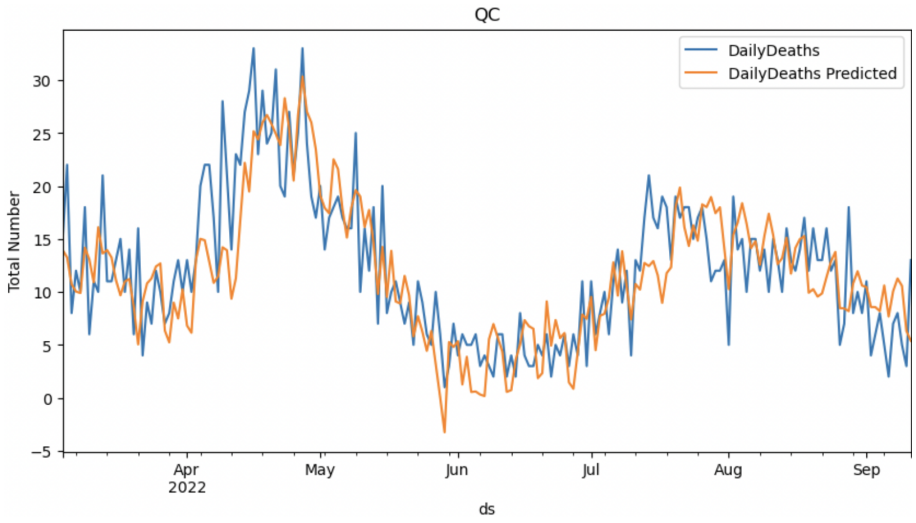


Fig. 6.4.31: Prediction Of Daily Deaths Deep Learning Time Series Framework Test Time Line (Region Quebec)

respective data points.

In Fig 6.4.31 We have Test Time line Prediction view of all the points and corresponding Daily Deaths for Quebec Province. This is prediction done by our Framework using LSTMs after the same was trained on respective data points.

6.4.6 Time Series Model Results Daily Hospitalisations

Results for Daily Hospitalisations are discussed below

In Fig 6.4.32 We have Full Time line Prediction view of all the points and corresponding Daily Hospitalisations for Quebec Province. This is prediction done by our Framework when after the same was trained and validated on respective data points.

In Fig 6.4.33 We have Test Time line Prediction view of all the points and corresponding Daily Hospitalisations for Quebec Province. This is prediction done by our Framework using LSTMs after the same was trained on respective data points.

In Fig 6.4.34 We have Test Time line Prediction view of all the points and corresponding Daily Hospitalisations for Quebec Province. This is prediction done by our Framework using Prophet after the same was trained on respective data points.

In Fig 6.4.35 We have Full Time line Prediction Percentile view of all the points

Time Series Models	**R^2**	**RMSE**	**MAE**	**MaxAE**	**[1455.26] + Training Time**
ARIMA	-2.60	12.9305	10.6060	25.6602	**1.62**
DL-TS Framework-Prophet	0.86	5.8060	4.1024	41.3995	2.46
DL-TS Framework-LSTMs	**0.94**	**4.2593**	**3.4009**	**13.7902**	1553.32
DL-TS Framework-DeerAREstimator	0.90	4.9221	3.5213	16.5223	348.46

Table 6.4.9: Predicting Daily Dealths (Region Quebec)

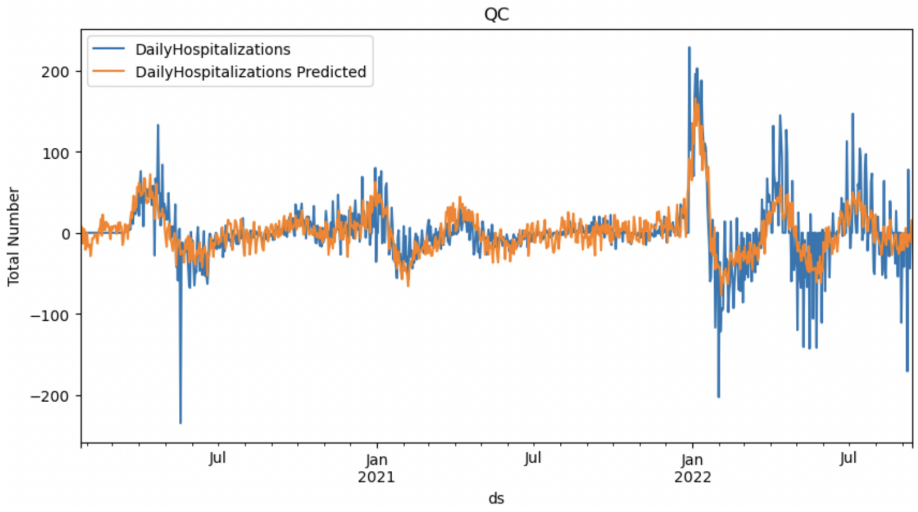


Fig. 6.4.32: Prediction Of Daily Hospitalisations Deep Learning Time Series Framework Full Time Line (Region Quebec)

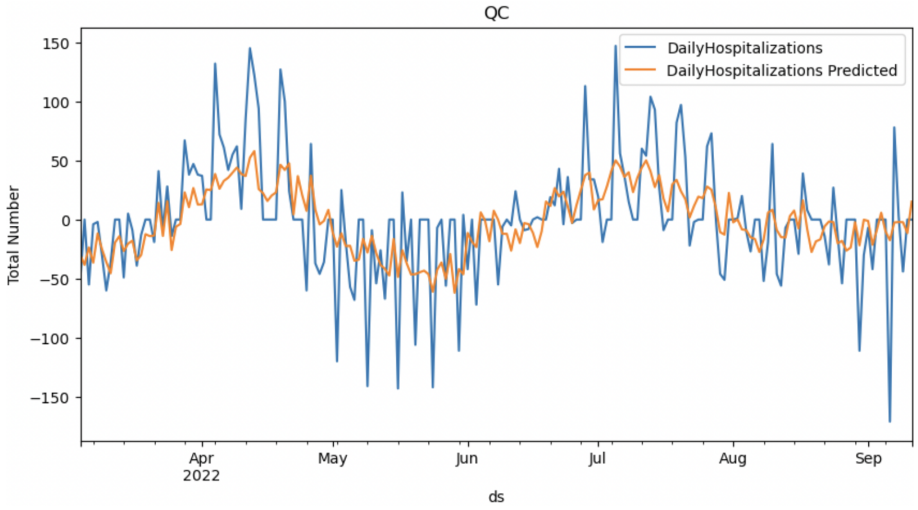


Fig. 6.4.33: Prediction Of Daily Hospitalisations Deep Learning Time Series Framework Test Time Line (Region Quebec)

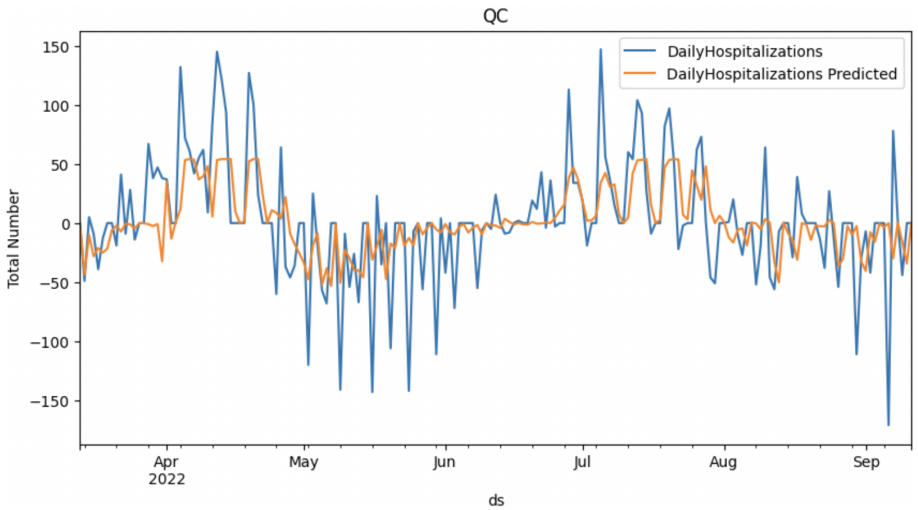


Fig. 6.4.34: Prediction Of Daily Hospitalisations Deep Learning Time Series Framework Test Time Line (Region Quebec)

Time Series Models	**R^2**	**RMSE**	**MAE**	**MaxAE**	**[1539.64] + Training Time**
ARIMA	-0.61	62.7439	50.4220	185.9979	**1.89**
DL-TS Framework-Prophet	0.80	37.4074	28.5497	208.0080	2.30
DL-TS Framework-LSTMs	**0.87**	**26.0494**	**17.7143**	**153.4828**	1505.32
DL-TS Framework-DeerAREstimator	0.82	41.0655	27.8959	170.8992	342.46

Table 6.4.10: Predicting Daily Hospitalisations (Region Quebec)

and corresponding Daily Hospitalisations for Ontario Province. This is prediction done by our Framework when after the same was trained and validated on respective data points.

In Fig 6.4.36 We have Test Time line Prediction view of all the points and corresponding Daily Hospitalisations for Ontario Province. This is prediction done by our Framework after the same was trained on respective data points.

6.5 Discussions

The results showcased in the preceding section underscore the remarkable effectiveness of our Regression Framework in predicting critical COVID-19 epidemic trends, encompassing total daily cases, deaths, and hospitalizations. When compared to conventional and contemporary regression models, it emerged as the superior performer, excelling across a spectrum of key evaluation metrics, including R-squared (R^2), Mean Absolute Error (MAE), Max Absolute Error (MaxAE), and Root Mean Squared Error (RMSE). This achievement was underpinned by the framework's capacity to furnish a scalable and sustainable model architecture, which exhibited exceptional proficiency in capturing a broad spectrum of trends.

However, as the focal point shifted towards the intricate dynamics governing daily

Time Series Models	**R^2**	**RMSE**	**MAE**	**MaxAE**	**[1539.64] + Training Time**
ARIMA	-0.03	194.2205	109.8348	1394.4811	**1.89**
DL-TS Framework-Prophet	0.80	147.3608	77.9118	1243.3523	2.30
DL-TS Framework-LSTMs	**0.85**	**89.1296**	**52.4412**	**1158.5178**	1505.32
DL-TS Framework-DeerAREstimator	0.81	109.3242	67.3422	1200.3536	342.46

Table 6.4.11: Predicting Daily Hospitalisations (Region Ontario)

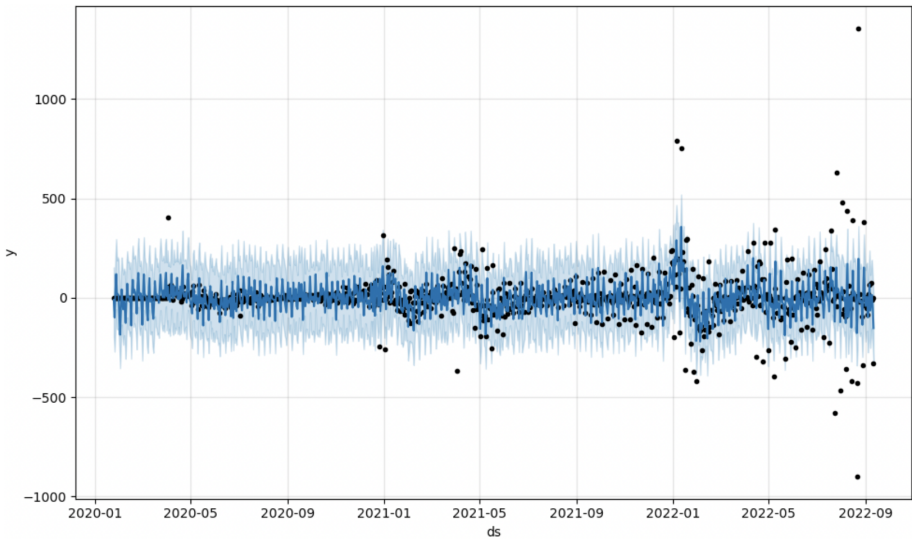


Fig. 6.4.35: Prediction Of Daily Hospitalisations Deep Learning Time Series Framework Test Time Line (Region Ontario)

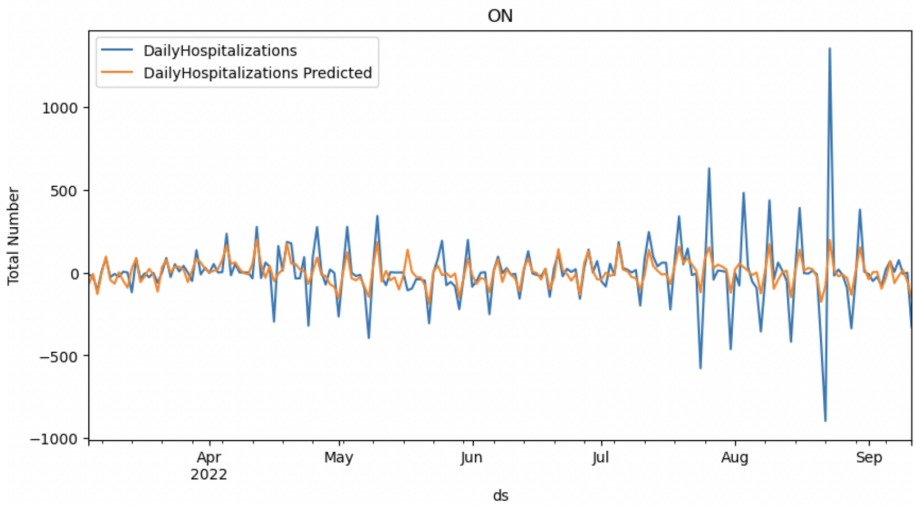


Fig. 6.4.36: Prediction Of Daily Hospitalisations Deep Learning Time Series Framework Test Time Line (Region Ontario)

cases, deaths, and hospitalizations, the Regression Framework encountered its limitations. Enter our specialized Deep Learning Time Series Framework, tailored to tackle the intricate challenges of predicting these daily trends. The Regression Framework assumed a pivotal role in this transition, functioning as the foundational element upon which our Deep Learning Time Series Framework was constructed.

In a comprehensive analysis of Time Series Models within our framework, their capabilities and trade-offs have come to light. These models consistently outperformed traditional statistical models, excelling in metrics such as R-squared, Mean Absolute Error, Max Absolute Error, and Root Mean Squared Error. However, they incurred a trade-off in training time, especially in the case of LSTMs. Despite the additional time investment during training, the enhanced prediction accuracy more than justifies this drawback. In terms of maintainability and scalability, the complexity of individual models for each region hindered the practicality of Prophet and LSTMs. In contrast, DeepAREstimator, offering comparable performance to LSTMs with significantly reduced training time, emerged as the most practical and sustainable choice. This comprehensive approach addresses multiple facets of epidemic trend prediction, making it an ideal solution for real-world applications where prediction performance, training time, and maintainability are essential considerations.

To sum up, our research emphasizes the impressive performance of Time Series Models within our framework, surpassing traditional models in prediction accuracy. Training time is a factor to consider, but the trade-off is justified by the improved predictions. DeepAREstimator, a single-model solution, offers a balance between performance, training time, and sustainability, making it the optimal choice for practical and reliable epidemic trend forecasting.

6.5.1 Statistical Stability and Reliability of Framework Results

The results achieved by both our Regression Framework and Deep Learning Time Series Framework not only showcase their effectiveness but also reflect an exceptional degree of statistical stability. A detailed analysis of these results over five repeated experiments has revealed an almost negligible variance, bordering on zero. The standard deviation, which quantifies the extent of variability in these results, is significantly smaller relative to the mean or expected values. This statistical insight into our models' performance indicates a level of stability and reliability that is truly remarkable.

Furthermore, it is noteworthy that both frameworks yield results with a standard deviation that is only a small fraction of the mean value. This observation suggests that the model's predictions consistently hover closely around the expected outcome. Such a high level of consistency enhances the frameworks' trustworthiness and reaffirms their potential to deliver dependable insights and predictions. The stability and reliability demonstrated by these frameworks, with minimal variance and standard deviation, reinforce their ability to provide valuable and consistent results.

6.5.2 Assumptions of Regression and Deep Learning Time Series Framework

- **Applicable Dataset** Our frameworks can only be applied when the dataset has all the features and properties of high cardinality multi variate multi-time

series dataset.

- **Stationary time series** Our frameworks assumed that the data is close to stationary, meaning that the statistical properties of the data, such as mean and variance, do not change over time. Certain preprocessing techniques, like differencing, can be used to achieve stationarity.

- **Assumption of Computational Resources:** Our frameworks assume that the computational resources and training times specified in the experimentation setup for each framework are available. The specified hardware, software, and computational infrastructure are assumed to be sufficient to train and run the models effectively. Additionally, it is assumed that the available resources can accommodate the computational demands of the deep learning models and statistical approaches utilized in the frameworks. The results are contingent on the availability and suitability of these computational resources.

6.5.3 Limitations of Regression and Deep Learning Time Series Framework

- **Data Requirements:** Our frameworks are limited in their applicability and effectiveness. They can only be applied when the dataset possesses all the features and properties of a high cardinality multivariate multi-time series dataset. If the dataset lacks the necessary characteristics, the frameworks may not perform optimally and may even fail to produce meaningful results.

- **Stationarity Assumption:** The frameworks operate under the assumption that the data is close to stationary, which means that the statistical properties of the data, such as mean and variance, do not change significantly over time. While certain preprocessing techniques like differencing can be employed to achieve stationarity, the performance of the frameworks may be compromised when applied to highly non-stationary data. This limitation affects

- **Computational Resources:** A critical limitation of our frameworks lies in

their reliance on specific computational resources. They assume the availability of the computational infrastructure, hardware, and software specified in the experimentation setup. If these resources are not accessible or inadequate, the frameworks' performance may be compromised. Furthermore, the effectiveness of the deep learning models and statistical approaches depends on the suitability and capacity of the available computational resources. This limitation may restrict the practicality of applying the frameworks in environments with limited computational capabilities.

- **Model Complexity:** The frameworks, particularly the Deep Learning Time Series Framework, can involve complex models with multiple layers and hyperparameters. This complexity can make them challenging to fine-tune and may require significant computational resources. Consequently, there is a limitation in their ease of use, especially for users without expertise in deep learning or machine learning.

- **Transferability:** While the frameworks are designed to predict COVID-19 epidemic trends, their transferability to other domains or types of datasets may be limited. They are tailored to the unique characteristics of epidemiological data and may not generalize well to different contexts, restricting their versatility.

- **Dependency on Training Data:** The frameworks rely on historical data for training. If the dataset does not cover a sufficiently long period, especially for time series data, the models may not capture long-term trends or exhibit less robust performance. This limitation is particularly relevant for forecasting epidemics, where historical context is crucial.

- **Resource Intensiveness:** Both frameworks can be computationally intensive, especially for training deep learning models. The need for substantial computational resources, including powerful GPUs and time, makes them less accessible in resource-constrained settings.

6.5.4 Contributions

- **A Comprehensive COVID-19 Canadian Dataset:** One of the foremost contributions of this research is the development and curation of a comprehensive COVID-19 Canadian dataset. This dataset serves as a foundational resource for understanding the dynamics of the pandemic across different regions of Canada. It offers a rich collection of data, encompassing various aspects of the epidemic, including environmental factors, government actions, medical parameters, mobility trends, and socio-economic conditions. The availability of such a dataset is invaluable for researchers, policymakers, and healthcare practitioners seeking to gain insights into the pandemic's multifaceted impact. This dataset, made accessible to the scientific community, is a lasting contribution to the fight against COVID-19 and future epidemic research.

- **Frameworks for Multivariate Multi-Timeseries Dataset:** The development of two distinct frameworks tailored to address the complexities of multivariate multi-timeseries datasets is a pivotal contribution of this research. These frameworks, specifically the Regression Framework and Deep Learning Time Series Framework, have been meticulously designed to cater to the intricacies of COVID-19 epidemic trend prediction. They provide researchers and practitioners with versatile tools for analyzing and forecasting various epidemic trends, ranging from total cases and deaths to daily fluctuations. These frameworks extend their utility beyond epidemiology and hold promise for addressing analogous challenges in other domains characterized by high-cardinality, multivariate, and multi-timeseries datasets.

- **Comprehensive Single AI Model for Epidemic Trends:** Another notable contribution is the development of a comprehensive single AI model capable of predicting a diverse range of epidemic trends. In a field where the conventional approach involves using distinct models for each type of prediction, this unified model offers both efficiency and scalability. The AI model, as a foundational element in the frameworks, exhibits robust performance across different

trends, making it an attractive solution for applications requiring multifaceted analysis of epidemic data. This singular model streamlines the modeling process, enhancing maintainability and reducing complexity in the development of predictive tools.

- **Critical Feature Identification:** A critical aspect of this research is the identification of key features that significantly influence the prediction of epidemic trends. The frameworks systematically extract relevant features from the dataset and pinpoint those that exhibit a strong correlation with the target variables. The ability to discern these critical features enhances our understanding of the underlying dynamics driving the epidemic. Moreover, it empowers researchers and policymakers to prioritize interventions and allocate resources effectively. By isolating these influential variables, this research simplifies the complexity of multivariate data and equips decision-makers with actionable insights.

CHAPTER 7

Conclusion and Future Work

7.1 Regression Framework

The Regression Framework has proven its mettle as a formidable tool in the domain of time series regression prediction. It has showcased its superiority over traditional and state-of-the-art regression models, excelling in a diverse range of critical metrics, including R-squared (R^2), Mean Absolute Error (MAE), Max Absolute Error (MaxAE), and Root Mean Squared Error (RMSE). This distinguished performance has underscored the Regression Framework's efficacy in deciphering complex dynamics and forecasting pivotal epidemic trends such as total daily cases, total daily deaths, and total daily hospitalizations.

Nonetheless, the Framework does have one notable limitation—prolonged training times. In comparison to traditional models, its computational demands mean a relatively lengthier training period. While its predictive capabilities are noteworthy, the longer training times may prove impractical in scenarios requiring real-time decision-making or rapid response.

Despite this drawback, the Framework has managed to accomplish an extraordinary feat: the establishment of a unified model architecture that efficiently predicts a broad spectrum of epidemic trends. This scalability and maintainability are pivotal achievements, streamlining model management and reducing complexity and resource requirements.

However, the Regression Framework encounters certain challenges in specific scenarios, particularly when tasked with predicting daily cases, daily deaths, and daily

hospitalizations. It struggles to establish a baseline for predicting these daily trends, as regional features and other attributes in the dataset fail to capture the intricacies of these fluctuations. Additionally, the distribution of these daily trends often exhibits similarities across multiple regions, adding an extra layer of complexity to the prediction task.

To address these limitations and delve deeper into the dynamics of daily trends, a specialized approach becomes imperative. Thus, our Deep Learning Time Series Framework was introduced, tailored to handle the nuances of predicting daily cases, deaths, and hospitalizations. The Regression Framework plays a pivotal role in this transition, serving as the foundational component. It contributes a vital element: the initial encoding representation for each date and its associated features. These encodings, meticulously generated by a deep neural network, serve as valuable input for our time series models, including Prophet, Long Short-Term Memory networks (LSTMs), and the potent DeepAREstimator.

This dual-pronged approach allows us to harness the strengths of both regression and time series modeling, optimizing our prediction capabilities and enhancing the depth of our insights. It offers a comprehensive solution that addresses the challenges posed by the diverse nature of epidemic trend data. Consequently, the Regression Framework's pivotal role as a foundational element ensures the seamless transition into a more specialized time series modeling framework, empowering us to unravel the complex dynamics of daily cases, deaths, and hospitalizations with finesse and precision.

7.2 Deep Learning Time Series Framework

The evaluation of Time Series Models within our framework has yielded remarkable insights into their capabilities and trade-offs. The performance of these models was benchmarked against traditional statistical models, revealing compelling results.

7.2.1 Outperformance in Prediction Metrics

The Time Series Models, including Prophet and LSTMs, showcased superior performance when compared to traditional statistical models. Across crucial evaluation metrics such as R-squared (R^2), Mean Absolute Error (MAE), Max Absolute Error (MaxAE), and Root Mean Squared Error (RMSE), the Time Series Models consistently outperformed their statistical counterparts. This marked improvement in prediction accuracy emphasizes the power of these models in capturing the intricate dynamics of epidemic trends.

7.2.2 Trade-off: Training Time

While the Time Series Models demonstrated excellence in prediction, a notable trade-off emerged in terms of training time. These models, particularly LSTMs, required more time to train compared to statistical approaches like ARIMA. The increased training duration is a noteworthy consideration, especially in scenarios where timely predictions are imperative. However, it is essential to recognize that the benefit of enhanced prediction accuracy often outweighs the incremental time investment during training.

7.2.3 Maintainability and Scalability

In the pursuit of models that not only perform well but are also practical for real-world applications, the concept of maintainability and scalability takes precedence. Prophet and LSTMs, although effective, pose challenges in this regard. Each region necessitates its own model, rendering these approaches less sustainable and maintainable. This inherent complexity can be a hindrance, particularly in situations demanding consistent and efficient model management.

7.2.4 The DeepAREstimator Advantage

To address the challenges of training time, sustainability, and maintainability, our Time Series Framework offers the promising solution of DeepAREstimator. While it

may not exhibit the same peak performance as LSTMs, it offers a performance that is comparable or slightly inferior. This marginal difference in performance is counterbalanced by a significantly reduced training time. DeepAREstimator also excels in practicality, as it requires only a single model for all regions, ensuring sustainability and maintainability. This approach presents an effective compromise, offering good performance without overwhelming computational demands.

7.2.5 Selecting the Optimal Solution

The choice of the optimal model and framework combination depends on specific priorities. If time is the sole concern, the framework with the Prophet model emerges as the best option, offering timely predictions with decent performance. However, for scenarios where prediction performance takes precedence, the framework with LSTMs proves to be the most suitable choice. It exhibits the best results across multiple metrics but entails longer training times. When a balanced solution is sought, combining good performance with decent training times and sustainable, maintainable models, the framework with DeepAREstimator stands as the ideal choice. It presents a holistic approach, effectively addressing multiple aspects of epidemic trend prediction.

In conclusion, our research reaffirms that Time Series Models, embedded within our framework, significantly outperform traditional statistical models across various evaluation metrics. The training time required for these models can be a limitation, but this is often justified by the substantial improvement in prediction accuracy. While Prophet, LSTMs, and DeepAREstimator each offer unique advantages, the latter emerges as a well-rounded choice. It presents an effective blend of performance, training time, sustainability, and maintainability, making it a practical option for governments and organizations seeking reliable epidemic trend predictions.

7.3 Future Works

Future work in this domain offers a multitude of exciting opportunities to enhance our current frameworks and extend their application.

- Firstly, expanding the dataset to encompass not just Canadian provinces but also those of other countries would open the door to international-level predictions. This cross-border dataset could enable us to develop models capable of forecasting epidemic trends on a global scale, offering valuable insights into the dynamics of pandemics across countries and continents.

- Secondly, there's ample room for experimentation within our feature selection, feature extraction, and model selection processes. Exploring alternative techniques and methods could potentially yield even more accurate and efficient frameworks. This avenue of research offers the potential for fine-tuning and optimizing our existing models to further improve their predictive performance.

- Thirdly, the problem at hand can be transformed into various structural representations, including knowledge graphs, paving the way for the application of sophisticated methods from the domain of graph theory and network analysis. In this context, techniques such as graph embeddings and community detection can be harnessed to uncover hidden patterns and relationships within the epidemic data. The incorporation of semantic web technologies, linked data, and ontologies could facilitate a more comprehensive understanding of the underlying factors affecting epidemic trends. By adopting a knowledge graph framework, we can explore the use of graph-based algorithms like network centrality measures to identify critical nodes and key influencers in the context of pandemic dynamics.

- Additionally, Natural Language Processing (NLP) techniques can be employed to extract valuable insights from textual data sources, augmenting our ability to predict and respond to epidemic trends effectively. This expanded approach, leveraging the power of knowledge graphs and graph-based analytics, promises to provide a more holistic and nuanced understanding of the intricate dynamics behind epidemics, ultimately leading to enhanced predictive models and decision support systems.

- Lastly, the application of more advanced machine learning techniques, such as convolutional neural networks (CNNs), has the potential to significantly enhance our frameworks. Incorporating comprehensive and exhaustive deep learning models could pave the way for groundbreaking advancements in predictive accuracy. These models, which have proven effective in various domains, can be tailored to capture complex relationships in epidemic data, potentially leading to more reliable and precise predictions.

7.4 Summary

In conclusion, the future of this research domain holds the promise of both broadening our datasets to a global scale and deepening our methodologies by exploring alternative techniques and approaches. The application of knowledge graphs and advanced machine learning models, such as CNNs, can open new horizons for epidemic trend prediction, ushering in an era of more accurate, efficient, and globally applicable forecasting models.

Leveraging Natural Language Processing (NLP) techniques, particularly the decoder transformer architecture, presents an opportunity for substantial performance improvement. This is especially noteworthy when considering the availability of additional data for training purposes. Transformer techniques, as demonstrated in various domains, have consistently outperformed traditional models such as LSTMs and RNNs. The inherent capability of transformers to capture intricate patterns and relationships in data, coupled with their parallel processing capabilities, positions them as a promising choice for tasks requiring nuanced understanding and predictive accuracy. The potential for further enhancement becomes particularly pronounced when larger and diverse datasets are employed for training these transformer models. Therefore, the prospect of refining performance through the strategic integration of transformer architectures, specifically decoder transformers, becomes even more compelling with the prospect of gathering additional data for training.

Applying this framework to data from diverse domains holds the promise of en-

hancing its robustness. By extending the analysis to multiple domains, the framework can uncover patterns and insights that transcend specific contexts, providing a more comprehensive understanding of its applicability. This cross-domain exploration not only contributes to the framework's versatility but also offers valuable insights that can inform adjustments and refinements, making the framework more adaptable across a broader spectrum of domains. The varied challenges and nuances encountered across different domains serve as invaluable inputs, guiding the identification of potential improvements and tweaks to further optimize the framework's performance and applicability. Thus, a multi-domain approach not only fortifies the robustness of the framework but also serves as a catalyst for continuous refinement and evolution.

APPENDIX A

The Novel Canadian Dataset developed for this thesis is available on GitHub at the following link:

`https://github.com/swastikbagga03/Thesis`

REFERENCES

[1] Artificial Intelligence Approach to Predict the COVID-19 Patient's Recovery — doi.org. https://doi.org/10.1007/978-3-030-63307-3_8. [Accessed 06-12-2023].

[2] Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model - Applied Intelligence — doi.org. https://doi.org/10.1007/s10489-020-01942-7. [Accessed 06-12-2023].

[3] Machine learning based approaches for detecting COVID-19 using clinical text data - International Journal of Information Technology — doi.org. https://doi.org/10.1007/s41870-020-00495-9. [Accessed 06-12-2023].

[4] Machine learning-based CT radiomics method for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study — doi.org. https://doi.org/10.21037/atm-20-3026. [Accessed 06-12-2023].

[5] Redirecting — doi.org. https://doi.org/10.1016/j.jcv.2020.104431. [Accessed 06-12-2023].

[6] Redirecting — doi.org. https://doi.org/10.1016/j.asoc.2020.106580. [Accessed 06-12-2023].

[7] Redirecting — doi.org. https://doi.org/10.1016/j.chaos.2020.110050. [Accessed 06-12-2023].

[8] Redirecting — doi.org. https://doi.org/10.1016/j.chaos.2020.109864. [Accessed 06-12-2023].

[9] Redirecting — doi.org. https://doi.org/10.1016/j.chaos.2020.109850. [Accessed 06-12-2023].

[10] Redirecting — doi.org. https://doi.org/10.1016/j.chaos.2020.110086. [Accessed 06-12-2023].

[11] Redirecting — doi.org. https://doi.org/10.1016/j.chaos.2020.109853. [Accessed 06-12-2023].

[12] Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests — doi.org. https://doi.org/10.3389/fcell.2020.00683. [Accessed 06-12-2023].

[13] Canadian Institute for Health Information (2023). Canadian institute for health information (cihi).

[14] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

[15] COVID-19 Canada Open Data Working Group (2023). Covid-19 timeline canada.

[16] Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534.

[17] Environment and Climate Change Canada (2023). Environment and climate change canada.

[18] Google (2023). Google covid-19 mobility reports.

[19] Ho, S. and Xie, M. (1998). The use of ARIMA models for reliability forecasting and analysis. *Computers &amp Industrial Engineering*, 35(1-2):213–216.

[20] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[21] Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

[22] li Lu, J., zhen Wang, L., jia Lu, J., and yue Sun, Q. (2008). Research and application on KNN method based on cluster before classification. In *2008 International Conference on Machine Learning and Cybernetics*. IEEE.

[23] Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning.

[24] Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg.

[25] Meraihi, Y., Gabis, A. B., Mirjalili, S., Ramdane-Cherif, A., and Alsaadi, F. E. (2022). Machine learning-based research for COVID-19 detection, diagnosis, and prediction: A survey. *SN Computer Science*, 3(4).

[26] Molokwu, B. C., Shuvo, S. B., Kobti, Z., and Snowdon, A. (2021). A multi-task learning framework for covid-19 monitoring and prediction of ppe demand in community health centres.

[27] Public Health Agency of Canada (2023). 2019 novel coronavirus infection (covid-19).

[28] R-bloggers (2015). Centering and standardizing: Don't confuse your rows with your columns.

[29] Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191.

[30] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

[31] Shuvo, S. B., Bagga, S., and Kobti, Z. (2023). Covid-19 analysis in canada using deep learning and multi-factor data-driven approach with a novel dataset. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–4.

[32] Statista (2023). Statista.

[33] Statistics Canada (2023). Statistics canada.

[34] Sulaiman, M. A. and Labadin, J. (2016). Improved feature selection based on mutual information for regression tasks. *Journal of IT in Asia*, 6(1):11–24.

[35] Tabrizchi, H., Mosavi, A., Szabo-Gali, A., Felde, I., and Nadai, L. (2020). Rapid covid-19 diagnosis using deep learning of the computerized tomography scans. In *2020 IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE)*, pages 000173–000178.

[36] Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.

[37] The Global Economy (2023). The global economy.

VITA AUCTORIS

NAME:	Swastik Bagga
PLACE OF BIRTH:	Delhi, India
YEAR OF BIRTH:	1995
EDUCATION:	Masters in Computer Science - Artificial Intelligence Stream
	University of Windsor, M.Sc in Computer Science, Windsor, Ontario, 2023