231654

OFFICE OF THE PRESIDENT
RECEIVED
Date: JUL 18 2023
Time: 6:20 pm
By:

SLSU OFFICE OF RESEARCH SERVICES

Office of
Research Services

18 July 2023

Dr. DORACIE B. ZOLETA-NANTES
University President

Thru: **Dr. MARISSA C. ESPERAL**
Vice President for Research, Extension, Production,
Development and Innovation

Dear Mesdames,

The CHED-DBM Joint Circular 3, s. 2023 (commonly known as the new instrument for Faculty Re-Classification) requires that Peer Reviewer engagement of faculty members in academic journals receive proper authorization from the President or the concerned Vice President. However, these guidelines were issued towards the end of the coverage period of the 1st Cycle of the Joint Circular (July 1, 2019–July 31, 2023).

As you may be aware, peer review requests from academic journals are normally directly communicated by editors to the peer reviewer and not through the institution where s/he may be affiliated. In consultation with the Institutional Evaluation Committee, I was informed that the CHED provides leeway for additional evidence for Peer Reviewer engagement – that a list of institutionally-recognized peer reviewer engagement would be enough as additional evidence for this cycle.

In this regard, I wish to respectfully seek your **approval in principle** of the participation of faculty members listed in the attached file. Rest assured that the ORS thoroughly screened these reported Peer Reviewer engagement of our faculty members to include only those done with reputable journal publications and book publishers.

We look forward to your usual support on this matter as this will contribute greatly to the career development of our dedicated faculty researchers.

Thank you very much!

Very truly yours,

NICANOR L. GUINTO, PhD
Director, Office of Research Services

Recommending Approval:

MARISSA C. ESPERAL, PhD
Vice President for Research, Extension,
Production, Development and Innovation

APPROVED / DISAPPROVED

Doracie B. Zoleta-Nantes, PhD
University President

JUL 19 2023

SOUTHERN LUZON STATE UNIVERSITY
Office of Research Services

C E R T I F I C A T I O N

This is to certify that the **peer reviewer engagement** of the personnel named below are approved in principle as they have been invited to review journal articles and/or book proposals while being affiliated with the University. For having been directly contacted by Editors of reputable journals and book publishers, their recognized expertise and leadership in their respective areas of research specialization contributed significantly to building the good name of Southern Luzon State University in local and international academic circles.

Name	Academic Rank	College/ Campus	Area of Research Specialization	Journal Name/ Book Publisher that made the request	Coverage/ Readership	Indexed/ Published by:	Tentative Title of the Article/ Book Proposal reviewed	Date when the invitation is received:	Date when the review was sent back to the editor:
AGUDILLA, MARY ANN R.	ASSOCIATE PROFESSOR 4	CAG	BIODIVERSITY, INSECTS, ECOSYSTEM VALUATION	PHILIPPINE JOURNAL OF SCIENCE	International	Scopus	SETTING THE INITIAL CARBON TAX RATE FOR THE CARBON TAX POLICY IN THE PHILIPPINES THROUGH THE SOCIAL COSTS OF CARBON AND WILLINGNESS TO PAY METHODS, AND THE CORRESPONDING BENEFIT-COST ANALYSIS	12/11/2022	1/2/2023
AGUDILLA, MARY ANN R.	ASSOCIATE PROFESSOR 4	CAG	BIODIVERSITY, INSECTS, ECOSYSTEM VALUATION	ACADEMIA-BIOLOGY	International	Academia Publishing	TREE HEIGHT, CANOPY COVER AND LEAF LITTER PRODUCTION OF RHIZOPHORA APICULATA IN BAGANGA, DAVAO, ORIENTAL, PHILIPPINES	1/11/2023	1/27/2023
Alinea, Jess Mark L.	Assistant Professor I	Lucena Campus	TVET, Technical Teacher Education, Curriculum and Instruction	Journal of Technical Education and Training	International	Scopus	The Role of Al-Balqa Applied University in Developing Vocational Education in Jordan	10/26/2021	11/2/2021
Alinea, Jess Mark	Assistant Professor I	Lucena Campus	TVET, Technical Teacher Education,	Journal of Technical	International	Scopus	Training-based Assessment of Employees Performance: A Case Study of Bahir Dar	12/27/2021	1/5/2022

			Communication	Applied Linguistics					
Guinto, Nicanor L.	Associate Professor III	College of Arts and Sciences	Sociolinguistics, Discourse Analysis, Communication	rEFLections	International	Scopus/ King Mongkut's University, Thailand	Filipino Non-Native English-Speaking Teachers and the Bias in Their Own Backyard	07/10/2023	07/19/2023
Maaliw, Renato III R.	Associate Professor II	CEN	Computer Vision, Machine Learning, Data Analytics	Cogent Engineering	International	Scopus, Web of Science, ASEAN Citation Index	Integrating Video Feedback Into Architectural Design Education to Engage Diverse Learning Styles	3/27/2023	4/20/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Machine Learning, Computer Vision, Data Analytics	Healthcare Analytics (Elsevier)	International	Scopus, Web of Science, ASEAN Citation Index	Prediction of Systolic and Diastolic Blood Pressures Using Machine Learning	5/4/2023	5/16/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics	Engineering (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Using ARIMA to Predict the Growth in the Subscriber Data Usage	11/4/2022	11/14/2022
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Analytics	Sensors (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Missing Traffic Data Imputation with a Linear Model Based on Probabilistic Principal Component Analysis	12/2/2022	12/10/2022
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	Sensors (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Using Machine Learning on V2X Communications Data for VRU's Collisions Predictions	12/23/2022	12/26/2022
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics	Applied Science (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Performance Predictions of Sci-Fi Films via Machine Learning	1/31/2023	2/5/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	Sustainability (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Thermal Images Classifications of Solid Wastes with Deep Convolutional Neural Networks	2/15/2023	2/25/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	Sustainability (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Static Evaluation of a Midimew Connected Torus Network for Next Generation Supercomputers	3/2/2023	3/13/2023
Maaliw,	Associate	College of	Computer Vision,	Journal of	International	Scopus, Web of	Machine-Learning-Based Composition	3/23/2023	4/1/2023

Renato III R.	Professor II	Engineering	Machine Learning, Data Analytics, Computer Engineering	Nuclear Engineering (MDPI)		Science, ASEAN Citation Index	Analysis of the Stability of V–Cr–Ti Alloys		
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	Mathematics (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	A Federated Personal Mobility Service in Autonomous Transportation	5/19/2023	5/29/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	IJERPH (MDPI)	International	Scopus, Web of Science,	Machine Learning in Predicting Severe Acute Respiratory Infection	6/6/2023	6/11/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	Journal of Theoretical and Applied Electronic Commerce Research	International	Scopus, Web of Science, ASEAN Citation Index	Unveiling the Power of ARIMA, Support Vector Machine and Random Forest Regressors for the Future of Dutch Employment Market	6/14/2023	6/23/2023
Mabunga, Zoren P.	Instructor 1	College of Engineering	Artificial Intelligence, Electronics and Communication Engineering, Internet of Things	2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA 2022)	International	Scopus	Semi Autonomous Detection of Bite Points for a Surgical Needle	2/24/2022	3/7/2022
Mabunga, Zoren P.	Instructor 1	Engineering	Artificial Intelligence, Electronics and Communication Engineering, Internet of Things	IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC-2021)	International	Scopus	1. A Survey of Vulnerability Management Using Machine Learning Techniques, 2. An Adaptive Algorithm based on Interference Aware Cooperative Energy Efficiency Maximization for 5G UltraDense Networks, 3. GRAMIN GENIE-A SMART KIOSK, 4. An Automated Deep Learning Model for Detecting Sarcastic Comments,	7/2/2021	8/12/2021

YAO, CLAIRE ANN M.	ASSISTANT PROFESSOR IV	CABHA MAIN	BUSINESS ENTREPRENEURSHIP, PRODUCT DEVELOPMENT, TOURISM, LEISURE, AND HOSPITALITY	PATHWAY TO REFEREED JOURNAL PUBLICATION IN THE FIELD OF BUSINESS	Local	INSTITUTIONAL	PROBLEMS ENCOUNTERED BY MSME'S IN TAGUIG CITY AND THE ACTION TO COUNTER THE POSSIBLE EFFECTS OF ASEAN INTEGRATION: A SITUATION ANALYSIS	3/24/2020	4/4/2020
YAO, CLAIRE ANN M.	ASSISTANT PROFESSOR IV	CABHA MAIN	BUSINESS ENTREPRENEURSHIP, PRODUCT DEVELOPMENT, TOURISM, LEISURE, AND HOSPITALITY	PATHWAYS TO REFEREED JOURNAL IN THE FIELD OF BUSINESS	Local	INSTITUTIONAL	MANYAMAN MANGAN QUENI (DELICIOUS TO EAT HERE):SUCCESS FACTORS OF SELECTED RESTAURANT ENTREPRENEURS IN PAMPANGA	4/16/2020	4/21/2020

Issued this 19th day of July 2023 at Southern Luzon State University, Lucban, Quezon.

NICANOR L. GUINTO, Ph.D.
Director, Office of Research Services

MARISSA C. ESPERAL, Ph.D.
Vice President, REPDI

DORACIE B. ZOLETA-NANTES, Ph.D.
University President

RM

Renato Racelis Maaliw

(Maaliw III, Renato R. R.)

Southern Luzon State University

Web of Science ResearcherID: EXW-3524-2022

Published names	Maaliw, Renato R., III Maaliw, Renato R. Maaliw, Renato Racelis, III Maaliw Iii, R. R. Maaliw Iii, Renato R. R.
Published Organizations	Southern Luzon State University, Southern Luzon Univ
Subject Categories	Computer Science; Telecommunications; Engineering; Environmental Sciences & Ecology; Materials Science

Documents | Peer Review

Verified peer reviews

4	Sensors
3	Sustainability
2	Applied Sciences
2	Eng
2	Journal of Theoretical and Applied Electronic Commerce Research
1	Cogent Engineering
1	International Journal of Environmental Research and Public Health
1	International Journal of Retail and Distribution Management
1	Journal of Nuclear Engineering
1	Mathematics

Metrics

Open dashboard

Profile summary

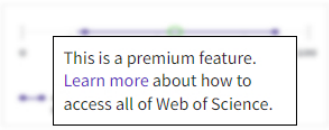
20	Total documents
15	Web of Science Core Collection publications
0	Preprints
18	Verified peer reviews
0	Verified editor records

Web of Science Core Collection metrics

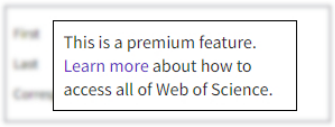
3 H-Index	**15** Publications in Web of Science
17 Sum of Times Cited	**10** Citing Articles
0 Sum of Times Cited by Patents	**0** Citing Patents

View citation report

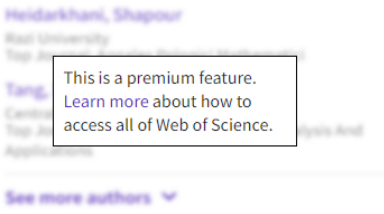
Author Impact Beamplot Summary



Author Position



You might be interested in...



See more authors

Co-authors

[Eng] Manuscript ID: eng-2039072 - Review Request

External Inbox ×

Eng Editorial Office <eng@mdpi.com>
to me, Eng, Simon
Fri, Nov 4, 2:30 PM (13 days ago)

Dear Dr. Maaliw,

We have received the following paper, submitted to Eng (https://www.mdpi.com/journal/eng/).

Type of manuscript: Article
Title: Using ARIMA to Predict the Growth in the Subscriber Data Usage
Special Issue: Feature Papers in Eng 2022
https://www.mdpi.com/journal/eng/special_issues/FP_in_Eng_2022

We kindly invite you to review this paper and evaluate its suitability for publication in Eng. The article abstract is available at the end of this message.

If you choose to accept this invitation, we would appreciate receiving your comments within 10 days. Please let us know if you are likely to need more time to complete your review.

Please click on the link below to let us know if you will be able to provide a review and access the full manuscript and review report form.

https://susy.mdpi.com/user/review/review/32477220/hxni3OwV

In recognition of the contribution of reviewers, for thorough and timely review reports we provide discount vouchers for Article Processing Charges (APCs) applicable for manuscripts accepted for publication after peer review in any MDPI journal. Advice for completing your review can be found at: https://www.mdpi.com/reviewers.

Please disclose any potential conflicts of interest you might have concerning the manuscript's contents or the authors.

If you are not able to review this manuscript, we kindly ask you to decline by clicking on the above link such that we can continue processing this submission. We would also appreciate any feedback you can provide at this time (i.e., your general impression regarding the quality of this manuscript) and any suggestions for alternative expert reviewers.

Eng is an open access journal of MDPI. Thank you very much for your consideration and we look forward to hearing from you.

Kind regards,
Mr. Simon Zhao
Assistant Editor
MDPI Eng Editorial Office
St. Alban-Anlage 66, 4052 Basel, Switzerland
E-Mail: eng@mdpi.com
http://www.mdpi.com/journal/eng

Manuscript details:
Journal: Eng
Manuscript ID: eng-2039072
Type of manuscript: Article
Title: Using ARIMA to Predict the Growth in the Subscriber Data Usage
Authors: Mike Nkongolo *, Jacobus Phillippus Van Deventer
Special Issue: Feature Papers in Eng 2022
https://www.mdpi.com/journal/eng/special_issues/FP_in_Eng_2022

Abstract: Telecommunication companies collect a deluge of subscriber data without retrieving substantial insights. The exploratory analysis of this type of data will facilitate the prediction of geographic, demographic, financial, and internet data which can be valuable to the decision-making process of telecommunication companies, but only if the retrieved insights have strategic plans and actions. The exploratory analysis of subscriber data was implemented in this research to predict subscriber usage trends based on historical time-stamped data. The predictive outcome was unknown but approximated using the data at hand. We have used 730 data points selected from the Insights Data Storage (IDS). These data points were collected from the hourly statistic traffic table and subjected to exploratory data analysis to predict the growth in subscriber data usage. The Auto-Regressive Integrated Moving Average (ARIMA) model was used to forecast. In addition, we used the normal Q-Q, correlogram, and standardized residual metrics to evaluate the model. This model showed a p-value of 0.007 with an accuracy of 90%. These results support our hypothesis predicting an increase in subscriber data growth. The compared results to the UGRansome dataset achieved the same accuracy values of 90% using the ARIMA model that predicted growth of 3 Mbps with a maximum data usage growth of 14 Gbps. In the experimentation, ARIMA was compared to the Convolutional Neural Network (CNN) and achieved the best results with the UGRansome data. These results provide

and achieved the best results with the UGRansome data. These results provide a road map for predicting subscriber data usage so that telecommunication companies can be more productive in improving their Quality of Experience (QoE). This study provides a better understanding of the seasonality involved in subscriber data usage's growth, exposing new network concerns and facilitating the development of novel predictive models.

Keywords: Time series forecasting, subscriber data, seasonality, ARIMA, telecommunication, UGRan- 21 some, QoE

Note: We discourage reviewers from recommending citation of their own work when not clearly necessary to improve the quality of the manuscript under review. Please state in your comments to the editor if you recommend citation of your own work and the reason for this recommendation.

MDPI partners with Publons (https://publons.com/in/mdpi/) to provide recognition for reviewers. Your credit will appear on Publons after a final decision on the paper and once you have claimed your review on the Publons website.

4 of 22

[Eng] Manuscript ID: eng-2039072 - Review Request Accepted

External Inbox ×

eng@mdpi.com
to me, Simon

Wed, Nov 9, 9:08 AM (8 days ago)

Dear Dr. Maaliw,

Thank you very much for agreeing to review this manuscript:

Manuscript ID: eng-2039072
Type of manuscript: Article
Title: Using ARIMA to Predict the Growth in the Subscriber Data Usage
Authors: Mike Nkongolo *, Jacobus Phillippus Van Deventer
Feature Papers in Eng 2022
https://www.mdpi.com/journal/eng/special_issues/FP_in_Eng_2022

The review report form can be found here:
https://susy.mdpi.com/user/review/review/32477220/hxni3OwV

The review report due date is: 16 November 2022

To ensure your anonymity throughout the peer review process, please do not include any identifying information in your review report either in the comments or in the metadata of any files that you upload. Please check the Guidelines for Reviewers: https://www.mdpi.com/reviewers

We look forward to receiving your valuable comments.

Kind regards,
Mr. Simon Zhao
Assistant Editor
MDPI Eng Editorial Office
St. Alban-Anlage 66, 4052 Basel, Switzerland
E-Mail: eng@mdpi.com
http://www.mdpi.com/journal/eng

*** This is an automatically generated email ***

Reply Reply all Forward

Review Report Form

Journal: Eng (ISSN 2673-4117)
Manuscript ID: eng-2039072
Type: Article

Title: Using ARIMA to Predict the Growth in the Subscriber Data Usage

Authors: Mike Nkongolo * , Jacobus Phillippus Van Deventer

Special Issue: Feature Papers in Eng 2022

Abstract: Telecommunication companies collect a deluge of subscriber data without retrieving substantial insights. The exploratory analysis of this type of data will facilitate the prediction of geographic, demographic, financial, and internet data which can be valuable to the decision-making process of telecommunication companies, but only if the retrieved insights have strategic plans and actions. The exploratory analysis of subscriber data was implemented in this research to predict subscriber usage trends based on historical time-stamped data. The predictive outcome was unknown but approximated using the data at hand. We have used 730 data points selected from the Insights Data Storage (IDS). These data points were collected from the hourly statistic traffic table and subjected to exploratory data analysis to predict the growth in subscriber data usage. The Auto-Regressive Integrated Moving Average (ARIMA) model was used to forecast. In addition, we used the normal Q-Q, correlogram, and standardized residual metrics to evaluate the model. This model showed a p-value of 0.007 with an accuracy of 90%. These results support our hypothesis predicting an increase in subscriber data growth. The compared results to the UGRansome dataset achieved the same accuracy values of 90% using the ARIMA model that predicted growth of 3 Mbps with a maximum data usage growth of 14 Gbps. In the experimentation, ARIMA was compared to the Convolutional Neural Network (CNN) and achieved the best results with the UGRansome data. These results provide a road map for predicting subscriber data usage so that telecommunication companies can be more productive in improving their Quality of Experience (QoE). This study provides a better understanding of the seasonality involved in subscriber data usage's growth, exposing new network concerns and facilitating the development of novel predictive models.

Review Report Form

Reviewer's Information (will not be revealed to authors)

Name: Dr. Renato Racelis Maaliw
Email: rmaaliw@slsu.edu.ph
Website: https://www.researchgate.net/profile/Renato-Maaliw-Iii
Affiliation: Southern Luzon State University

Research Keywords: big data; data mining; Machine Learning; Analytics; Computer Vision

Report 1 Hide Report and Author Response [-]

	High	Average	Low	No Answer
Originality / Novelty	( )	( )	(x)	( )
Significance of Content	( )	( )	(x)	( )
Quality of Presentation	( )	(x)	( )	( )
Scientific Soundness	( )	(x)	( )	( )
Interest to the readers	( )	( )	(x)	( )
Overall Merit	( )	( )	(x)	( )

Overall Recommendation

( ) Accept in present form
( ) Accept after minor revision (corrections to minor methodological errors and text editing)
(x) Reconsider after major revision (control missing in some experiments)
( ) Reject (article has serious flaws, additional experiments needed, research not conducted correctly)

English language and style

( ) English very difficult to understand/incomprehensible
( ) Extensive editing of English language and style required
(x) Moderate English changes required
( ) English language and style are fine/minor spell check required
( ) I don't feel qualified to judge about the English language and style

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	( )	(x)	( )	( )
Are all the cited references relevant to the research?	( )	(x)	( )	( )
Is the research design appropriate?	( )	( )	(x)	( )
Are the methods adequately described?	( )	( )	(x)	( )
Are the results clearly presented?	( )	(x)	( )	( )
Are the conclusions supported by the results?	( )	(x)	( )	( )

Comments and Suggestions for Authors

The ARIMA is the most common forecasting models, there are hundreds and even thousands of researches in forecasting using this one.

For me, I have to reject this paper due to novelty (nothing is new).

As a recommendation for improvements:

1. Explore different forecasting models, you can combine classical mathematical models and the strengths of neural networks specifically recurrent neural networks (RNN) using various ensemble algorithms

2. This paper does not explore the most important part of modeling, hyperparameter optimization, in my experience this is the most neglected part of of any forecasting models.

3. It will be also better to explore other factors that affects subscriber's data usage, specifically an multivariate forecasts.

4. This is a good paper but needs major revision on the methods as stated in number 1.

Less...

	Yes	No
Do you have any potential conflict of interest with regards to this paper?	( )	(x)
Did you detect plagiarism?	( )	(x)
Did you detect inappropriate self-citations by authors?	( )	(x)
Do you have any other ethical concerns about this study?	( )	(x)

1 of 22

[Eng] Manuscript ID: eng-2039072 - Acknowledgement - Review Received

External Inbox ×

eng@mdpi.com 10:16 AM (26 minutes ago)
to me, Eng, Simon

Dear Dr. Maaliw,

Thank you for submitting your review of the following manuscript:

Manuscript ID: eng-2039072
Title: Using ARIMA to Predict the Growth in the Subscriber Data Usage
Authors: Mike Nkongolo *, Jacobus Phillippus Van Deventer

Our Editorial Office and Academic Editors will contact you if they have any questions about your review report. We ask that you remain available, as far as possible, during the peer-review process in case of follow-up questions. To help us improve our services, we kindly ask you to fill in our online survey on the peer-review process at
https://www.surveymonkey.com/r/reviewerfeedbackmdpi

We encourage you to register an account on our submission system and bind your ORCID account (https://susy.mdpi.com/user/edit). You are able to deposit the review activity to your ORCID account manually via the below link:
https://susy.mdpi.com/user/reviewer/status/finished

We also invite you to contribute to Encyclopedia (https://encyclopedia.pub), a scholarly platform providing accurate information about the latest research results. You can adapt parts of your paper to provide valuable reference information for others in the field.

Kind regards,
Mr. Simon Zhao
Assistant Editor
MDPI Eng Editorial Office
St. Alban-Anlage 66, 4052 Basel, Switzerland
E-Mail: eng@mdpi.com
http://www.mdpi.com/journal/eng

*** This is an automatically generated email ***

Reply Reply all Forward

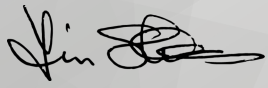
REVIEW CONFIRMATION CERTIFICATE

We are pleased to confirm that

Renato Racelis Maaliw

has reviewed 1 paper for the following MDPI Journal in 2022:

Engineering



Dr. Shu-Kun Lin, Publisher and President
Basel, 17 November 2022

MDPI is a publisher of open access, international, academic journals. We rely on active researchers, highly qualified in their field to provide review reports and support the editorial process. The criteria for selection of reviewers include: holding a doctoral degree or having an equivalent amount of research experience; a national or international reputation in the relevant field; and having made a significant contribution to the field, evidenced by peer-reviewed publications.

MDPI

Article

Using ARIMA to Predict the Growth in the Subscriber Data Usage

Mike Nkongolo [1]* and Jacobus Philippus van Deventer [1]

[1] Department of Informatics, Faculty of Engineering, Built Environment and Information Technology, University of Pretoria, South Africa; u21629545@tuks.co.za; phil.vandeventer@up.ac.za
* Correspondence: u21629545@tuks.co.za; mike@mavensysworx.com

Abstract: Telecommunication companies collect a deluge of subscriber data without retrieving substantial insights. The exploratory analysis of this type of data will facilitate the prediction of geographic, demographic, financial, and internet data which can be valuable to the decision-making process of telecommunication companies, but only if the retrieved insights have strategic plans and actions. The exploratory analysis of subscriber data was implemented in this research to predict subscriber usage trends based on historical time-stamped data. The predictive outcome was unknown but approximated using the data at hand. We have used 730 data points selected from the Insights Data Storage (IDS). These data points were collected from the hourly statistic traffic table and subjected to exploratory data analysis to predict the growth in subscriber data usage. The Auto-Regressive Integrated Moving Average (ARIMA) model was used to forecast. In addition, we used the normal Q-Q, correlogram, and standardized residual metrics to evaluate the model. This model showed a p-value of 0.007 with an accuracy of 90%. These results support our hypothesis predicting an increase in subscriber data growth. The compared results to the UGRansome dataset achieved the same accuracy values of 90% using the ARIMA model that predicted growth of 3 Mbps with a maximum data usage growth of 14 Gbps. In the experimentation, ARIMA was compared to the Convolutional Neural Network (CNN) and achieved the best results with the UGRansome data. These results provide a road map for predicting subscriber data usage so that telecommunication companies can be more productive in improving their Quality of Experience (QoE). This study provides a better understanding of the seasonality involved in subscriber data usage's growth, exposing new network concerns and facilitating the development of novel predictive models.

Keywords: Time series forecasting, subscriber data, seasonality, ARIMA, telecommunication, UGRansome, QoE

Citation: Mike Nkongolo and Jacobus Philippus van Deventer. Using ARIMA to Predict the Growth in the Subscriber Data Usage. *Journal Not Specified* **2022**, *1*, 0.

Received:
Accepted:
Published:

1. Introduction

The growth of competition in the telecommunications industry due to technological variety has facilitated the invention and expansion of new techniques for processing subscriber data to predict their behavior. Subscriber traffic represents all kinds of electronic data transmitted in the network. This data is usually in the form of network flows passing from one node to another [1]. We have used subscriber data collected from a network database and analyzed the patterns to predict the growth in subscriber data usage. The Network Subscriber Data Management (NSDM) approach is thus the relevant aspect of this research as it stands at the core network layer and stores valuable data used by various subscribers. The NSDM extracts subscribers' patterns from the Insights Data Storage (IDS) and monitors all real-time traffic of subscriber data [2]. We have used the NSDM module that considers subscriber data in a centralized and secure environment having a scalable repository named IDS (Figure 1).

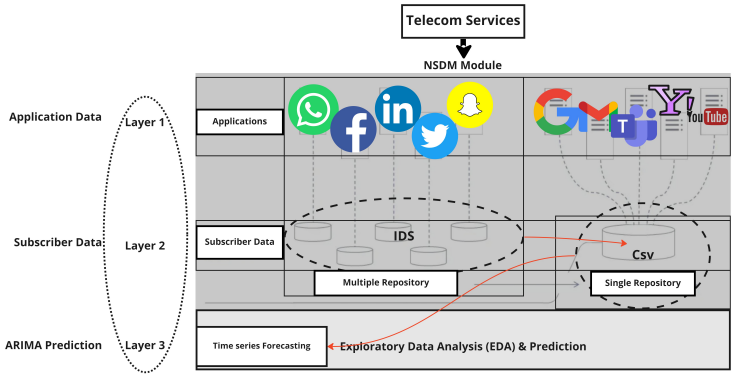


Figure 1. The NSDM architecture.

The IDS directory provides distributed and resilient subscriber patterns stored in a single repository. The ARIMA model was used on this repository to predict the growth in subscriber data usage (Figure 1). In this article, we describe the advantages of using seasonality to examine changes in subscriber data. A yearly repeating feature is known to exhibit seasonal properties. However, in this research, seasonality refers to repeated features of a fixed period in the subscriber dataset. In a time series analysis (TSA), if a pattern repeats at the same frequency or timestamp, it is seasonal [3]. Understanding seasonality in TSA can enhance the prediction performance of Machine Learning (ML) models. It can also assist in clearing the features by identifying the seasonality of time series samples and removing them from the original dataset. As a result, one can have a normalized dataset correlating input and output variables. The seasonality property can also provide more information about the seasonal component of the time series data that can provide insights to enhance predictors' performance [4].

Modeling seasonality ameliorates the data preparation and feature engineering steps. In each step, seasonal patterns can be extracted and modeled as input/output class labels with a supervised learning scheme. In adaptive computation, ARIMA is a class of predictive classifiers that provide linear outputs dependent on their previous observed values by using a combination of stochastic parameters [5]. We have selected an adequate time series forecasting model named ARIMA to predict subscriber data usage and analyze the seasonality, trends, and cycles of features. We have to consider seasonality as the time series data property used in the ARIMA model that implemented a distributed lag algorithm to forecast future subscriber data usage based on lagged parameters. This article implements a predictive ARIMA model using subscribers' data to study seasonality by predicting the growth in subscriber throughput.

Our research contribution. We propose the ARIMA model for subscriber data prediction using an unsupervised learning scheme. We have specifically implemented the ARIMA model with unlabelled features to predict the growth in subscriber data usage. In the model, the predictive layer forecasts the throughput rate fed into another layer that predicts the maximum usage growth. The remainder of the paper is structured as follows. Section 2 discusses the literature review and Section 3 the research methodology. Section 4 presents the ARIMA results and the comparative analysis using the UGRansome dataset. Section 5 presents future research directions and concluding remarks.

2. Literature Review

The literature surveys the predictive challenges of time series data with special attention on the ML model by highlighting the contributions of our proposed methodology.

2.1. Background

The authors in [6] used regression to learn the correlation between a time series and continuous variables.

The approach was to detect the correct coefficients to forecast various attributes. The regression model predicted annual rainfall using historical temperature values [7] with a Random Forest (RF) and Gradient Descent (GD) algorithm. The final results confirm the in-depth understanding of time series data to compute the optimal fitting algorithm. However, [8] attempted to predict respiratory rates using a sliding window that consists of three modules. The first module retrieved the signal of respiratory patterns; the second approximated the respiratory rates, and the third estimated the respiratory rate. A Gaussian-based regression process extracted the respiratory rate from various datasets. It also attempted to fit different Auto-Regressive (AR) algorithms to the retrieved signals. Unfortunately, the AR model failed to detect seasonality. In [9], Dynamic Time Warping (DTW) and K-Nearest Neighbor (KNN) used for time series forecasting exhibited a complexity time of 1-NN using the DTW that relied on the engineering of hand-crafted patterns. In [10], the Convolutional Neural Network (CNN) used on time-series data outperformed all other tested ML models. The author proposed a feature selection method to automate the learning from input variables. The learned patterns represent time series features with discriminatory layers. However, this technique relies on back-propagation that turns the Neural Network (NN) into an adequate feature selector.

According to [11], the juxtaposition of Recurrent Neural Networks (RNN) such as Long Short-Term Memory (LSTM) and CNN yielded enhanced accuracy for classification tasks with a range of 27% to 43% in comparison to other well-known ML models. The classification was also considered by [12] and assessed with J48, LSTM, RF, Support Vector Machine (SVM), and CNN. The LSTM-based-CNN outperformed other models with three hidden layers. In [13], the authors used regression to allocate company resources. In addition, the authors undertook a substantial review of well-known ML models for time series data forecasting, but [14] used the CNN to address multivariate time-series regression problems. The LSTM and Gated Recurrent Unit (GRU) portrayed transferable CNN units compared to other models. The research in [15] used LSTM with additional convolutional layers. The results provide a boost in predicting performance. Lastly, three CNN and four LSTM were implemented by [16] with an improved CNN execution time. Generally speaking, regression models using CNN and LSTM are the most optimal ML techniques used in the literature for time series data forecasting (Table 1).

Table 1. Comparative analysis

Source	Model	Limitation
[6]	RF & GD	Data understanding
[8]	Auto Regressive	Seasonality
[13]	CNN	Seasonality
[9]	DTW & KNN	Feature engineering
[10]	CNN	Back propagation
[11]	RNN & LSTM	Classification
[12]	LSTM & CNN	Classification
[15]	LSTM	Feature engineering
[16]	LSTM & CNN	Execution time
[14]	CNN	Biases

The limitation of the discussed research relies on dataset misunderstanding, lack of feature engineering, non-seasonal patterns, computational biases, and time complexity. Classifiers such as SVM and Decision Trees (DT) are also prone to error in terms of time series pattern prediction since they are not a better choice for forecasting (Table 1).

2.2. Time Series Data Analysis

Some attempts allow efficient computation of large-scale time series data. For instance, [17] implemented a Hadoop-based framework for accurate preprocessing of data which is important for feature selection.

Unlike [17], [18] concentrated on model selection by using MapReduce to compute the cross-validation that improved parallel rolling-window prediction using the training set of heterogeneous time series patterns. The predictive parameters computed the accuracy, but this technique could not tackle challenges associated with forecasting. In [19] and [20] multi-step forecasting was monitored by the ML models using the Spark environment. Specifically, [19] used H iterations to compute the multi-step prediction, while [20] implemented multivariate regression models using ML libraries. As a result, this technique was not scalable for forecasting. With this, one can use a sample of patterns instead of the original data to predict. For example, [21] provides an overview of forecasting big data using time series traffic. The paper provides a premise for time series data forecasting, but it is still complicated to implement the proposed techniques to deal with subscriber data and forecast the future. Some researchers investigate the underlying intuition of parallel computing models using time series data. Unfortunately, these models resulted in expensive computational time complexity.

For instance, [22] introduced a distributed approximator before the prediction calculation, requiring several iterations. Based on their frameworks, [23], [24] proposed recursive techniques with Bayesian prediction while [25] refined the estimator computation of quantile regression model through various rounds of classification. Another well-known methodology is the alternation of eigenvectors for convex optimization of time series data. This technique blends the seasonality of time series data with the convergence properties of predictors [26], but the streams complicate the forecasting prediction. We argue that a one-shot averaging computation is a straightforward technique to compute the prediction. This method requires only a single computational round [27]. Various studies used distributed learning that split features in a specific frequency domain where the time series patterns are used in the splitting process [28]. These algorithms model successive refinements with a limitation that requires re-implementing each estimator scheme, but slow in terms of convergence accuracy compared to existing predictors designed for time series data forecasting [29]. However, all mentioned articles in this section are crucial because they provide valuable recommendations regarding ML to forecast subscribers' usage data growth.

2.3. Mathematical Formulation of ARIMA

An Auto-Regressive Integrated Moving Average (ARIMA) model has a different moving average (MA), as well as auto-regressive (AR) components [30]. We use ARIMA(p, d, q) to denote an ARIMA model where the order of the AR module is (p, q) and d represents the number of differences needed for stationary series [30]. One can extend the ARIMA predictor to a seasonal ARIMA (SARIMA) model by incorporating additional seasonal patterns to handle time series properties that exhibit a strong seasonal characteristic [30]. We can use ARIMA(p, d, q)(P, D, Q) to formulate a SARIMA model. Here, the uppercase Q, P, and D denote the order of the AR model, the number required for seasonal/stationary series, and the MA order. Similarly, the seasonality period is denoted by m [31], [30]. An ARIMA(P, D, Q)$(p, d, q)_m$ model for time series $(y_t, t \in \mathcal{Z})$ has the following back-shift operator:

$$(1 - \sum_{i=1}^{p} \theta_i B^i) - (1 - \sum_{m=1}^{P} \alpha_i^m)(1 - B)^d (1 - B^m)^D y_t = (1 + \sum_{i=1}^{q} \gamma_i)(1 + \sum_{i=1}^{Q} \alpha_i B^m)\omega_t \quad (1)$$

Where B denotes the backward shift function, ω_t the white noise, m the seasonality length, θ, and α represent the AR parameters, γ, and ω refer to the seasonal parameters of the MA. This mathematical formulation represents two major combinations of seasonal parameters P, D, Q, and p, d, q, where:

- the number of auto-regressive terms is p,
- nonseasonal differences denoted by d,

- and the number of lagged predictive biases denoted by q.

The variation of these ARIMA parameters can identify the most optimal set of features in obtaining precise predictive values [32], [31], [30].

Cyclostationarity

In [33], analyzed cyclostationary properties of 0-day exploits were given. Boruta was the feature-based extraction method combined with Principal Components Analysis (PCA) to extract the most cyclostationary features from NSL-KDD, UGRansome, and KDD99 datasets. The RF and SVM were used to classify time series features. The supervised learning restricted the experiments, but our research implements an unsupervised learning scheme to study subscriber data usage prediction. Nevertheless, we have compared the ARIMA model performance applied to the UGRansome [34] and subscriber datasets to assess the forecasting performance of time-series data. The following section will present the research methodology, subscriber data, EDA, UGRansome dataset [35], and evaluation metrics.

3. Materials and Methods

3.1. Experimental Datasets

Figure 2 presents the research methodology where our framework provides subscriber data stored in the Insights Data Storage (IDS) module.

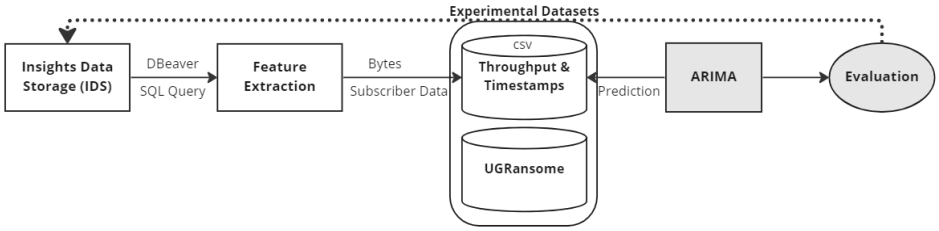
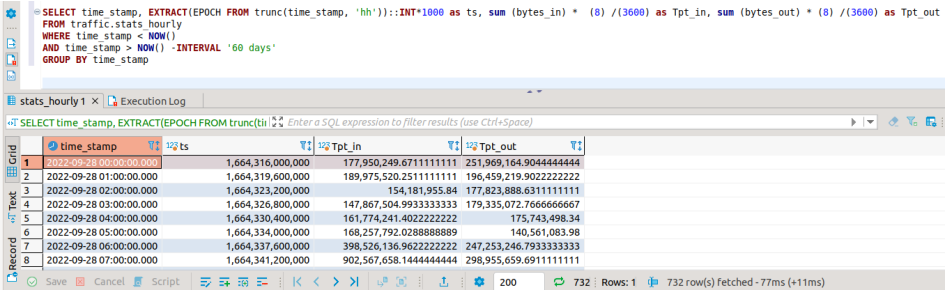


Figure 2. The experimental methodology.

The subscriber data was extracted from the real-time network traffic using a Structured Query Language (SQL). We pushed the features into a single comma-separated file and used EDA to visualize salient features of the network traffic. We have then obtained critical Key Performance Indicators (KPIs) that can support the prediction of data usage growth. The executed SQL retrieved the subscriber timestamps, incoming throughput, and outgoing throughout (Figure 3).



```
SELECT time_stamp, EXTRACT(EPOCH FROM trunc(time_stamp, 'hh'))::INT*1000 as ts, sum (bytes_in) *  (8) /(3600) as Tpt_in, sum (bytes_out) * (8) /(3600) as Tpt_out
FROM traffic.stats_hourly
WHERE time_stamp < NOW()
AND time_stamp > NOW() -INTERVAL '60 days'
GROUP BY time_stamp
```

stats_hourly 1 | Execution Log

	time_stamp	ts	Tpt_in	Tpt_out
1	2022-09-28 00:00:00.000	1,664,316,000,000	177,950,249.67111111111	251,969,164.9044444444
2	2022-09-28 01:00:00.000	1,664,319,600,000	189,975,520.25111111111	196,459,219.9022222222
3	2022-09-28 02:00:00.000	1,664,323,200,000	154,181,955.84	177,823,888.6311111111
4	2022-09-28 03:00:00.000	1,664,326,800,000	147,867,504.9933333333	179,335,072.7666666667
5	2022-09-28 04:00:00.000	1,664,330,400,000	161,774,241.4022222222	175,743,498.34
6	2022-09-28 05:00:00.000	1,664,334,000,000	168,257,792.0288888889	140,561,083.98
7	2022-09-28 06:00:00.000	1,664,337,600,000	398,526,136.9622222222	247,253,246.7933333333
8	2022-09-28 07:00:00.000	1,664,341,200,000	902,567,658.1444444444	298,955,659.6911111111

Save Cancel Script 200 732 Rows: 1 732 row(s) fetched - 77ms (+11ms)

Figure 3. The subscriber data.

The query extracts the timestamps (ts) by truncating them into a human-readable format (Year-Month-Time). The incoming throughput was computed using the following Equation 2:

$$Tpt_{in} = \frac{sum(bytes_{in}) * (8)}{36,000} \quad (2)$$

The SQL in Figure 3 illustrates this process. Equation 3 denotes the outgoing throughput computation:

$$Tpt_{out} = \frac{sum(bytes_{out}) * (8)}{36,000} \quad (3)$$

It is hourly-based statistics retrieved from the traffic stats table of the IDS for 60 days (Figure 3). In addition, we grouped results by timestamps. Retrieved patterns were converted into Comma-Separated Values (CSV) (Figure 4).

time_stamp	ts	Tpt_in	Tpt_out
2022-09-28 0:00:00	1664316000000	177950249.7	251969164.9
2022-09-28 1:00:00	1664319600000	189975520.3	196459219.9
2022-09-28 2:00:00	1664323200000	154181955.8	177823888.6
2022-09-28 3:00:00	1664326800000	147867505	179335072.8
2022-09-28 4:00:00	1664330400000	161774241.4	175743498.3
2022-09-28 5:00:00	1664334000000	168257792	140561084
2022-09-28 6:00:00	1664337600000	398526137	247253246.8
2022-09-28 7:00:00	1664341200000	902567658.1	298955659.7
2022-09-28 8:00:00	1664344800000	2042488462	612111154.3
2022-09-28 9:00:00	1664348400000	1833562607	639776109.2

Figure 4. The CSV format of the subscriber data.

The subscriber dataset has 730 entries with four attributes (human-readable timestamps, UNIX timestamps, incoming throughput (Tpt in), and outgoing throughput (Tpt out)). A timestamp represents the time when the subscriber traffic was collected [36]. The throughput is the flow that measures inputs/outputs movements within the network [36]. The following Figure 5 illustrates our research methodology.

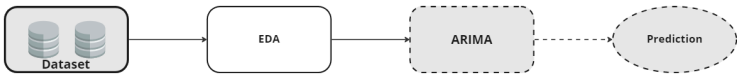


Figure 5. The research methodology.

The subscriber and UGRansome datasets are collected, then the EDA is executed before the computation of the ARIMA model that predicts the growth in subscriber data usage based on the current timestamp. The techniques discussed in the literature train ML classifiers with human-labeled features, but this supervised learning method uses limited samples. We have used an unsupervised learning technique whereby we did not label the features. The ARIMA model attempted to use data points $x_1 ... x_n$ and assigned predicted values $\Theta_1 ... \Theta_n$ using predefined parameters.

3.2. The UGRansome Characteristics

This dataset was created by extracting important features of two existing datasets (UGR'16 and ransomware) [35]. UGRansome is an anomaly detection dataset that includes normal and abnormal network activities [37]. The regular characteristic sequence makes up 41% of the dataset, whereas irregularity makes up 44%. The remaining 15% represents the predictive values of network attacks grouped into the signature (S), synthetic signature (SS), and anomalous (A) attacks (Fig. 6). Figure 6 depicts the signature attacks having a proportion of 44.02%, synthetic signature 28.71%, and anomaly 27.27%. A significant proportion of signature traffic means that the UGRansome threatening concerns are detectable. Regular threats, like User Datagram Protocol (UDP) and Botnet, provide about 9% for the anomalous category. The Internet Protocol (IP) and ransomware addresses have a ratio of 1% [33]. In addition, a ratio of two percent exists between communication protocols and ransomware addresses [35]. According to [35] and [33] the significant distribution of the UGRansome could be summed up in the following Figure 7. However, UGRansome is more redundant compared to subscriber data and we removed 28.2% of duplicate records during the feature extraction phase (Figures 8 and 7).

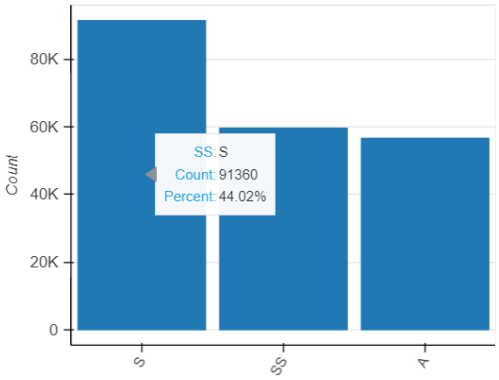


Figure 6. Distribution of network threats.

Dataset Statistics

Number of Variables	14
Number of Rows	207533
Missing Cells	0
Missing Cells (%)	0.0%
Duplicate Rows	58491
Duplicate Rows (%)	28.2%
Total Size in Memory	106.9 MB
Average Row Size in Memory	540.2 B
Variable Types	Numerical: 4 Categorical: 9 GeoGraphy: 1

Figure 7. The UGRansome data summary.

Dataset Statistics

Number of Variables	4
Number of Rows	730
Missing Cells	0
Missing Cells (%)	0.0%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	74.3 KB
Average Row Size in Memory	104.2 B
Variable Types	Categorical: 1 Numerical: 3

Figure 8. The subscriber data summary.

3.3. Exploratory Techniques

The exploratory analysis provides a set of techniques to understand the dataset. The results produced by the EDA can assist in mastering the data structure [38], as well as the distribution of the features, detection of outliers, and correlation within the dataset. Some of the statistical metrics used to evaluate the ARIMA model are standard deviation, correlation, mean, standardized residual, normal Q-Q, correlogram, theoretical quantile, p-value, and accuracy:

- **Standardized residual (r_i).** It measures the strength of actual and predicted values and indicates the significance of features [39] (r_i facilitates the recognition of features that contribute the most to the predictive values):

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{RSE\sqrt{1-h_i}} \tag{4}$$

  Where e_i is the iTh residual, RSE is the standard error of the residual model, and h_i the iTh leverage observation.
- **Normal Q-Q.** The normal Q-Q means normal Quantile-Quantile. It is a plot that compares actual and theoretical quantiles [39]. The metric considers the range of random variables to plot normal Q-Q using a probabilistic computation.

The x-axis represents the Z-score of the standardized normal distribution, but different formulations have been proposed in the literature to detect the plotting positions:

$$\frac{(k-a)}{(n+1-2a)}, \tag{5}$$

for some value between 0 and 1 [0,1]; which gives the following range (Equation 6):

$$\frac{K}{(n+1)} \Longleftarrow \frac{(k-1)}{(n-1)}. \tag{6}$$

- **Correlogram.** It is a correlational and statistical chart used in Time Series Analysis (TSA) to plot the auto-correlations sample r_h versus the timestamp lags h to check for randomness [39]. The correlation is zero when randomness is detected. Equation 7 denotes the auto-correlation parameter at h lag.

$$r_h = \frac{c_h}{c_0}, \tag{7}$$

where c_h is the auto-covariance coefficient and c_0 the variance function.

- **Augmented Dickey-Fuller (ADF) test.** This statistical metric tests the stationarity of time series data [39] by using a unit root metric that exists in a series of observations where $\alpha = 1$ as per the below Equation 8.

$$y_t \Longrightarrow \alpha_{t-1} + \beta x_e + \epsilon, \tag{8}$$

Here y_t represents the time series values at time t, but x_e is a separate time series variable.

- **Theoretical quantile.** The theoretical Q-Q explores the variable's deviation from theoretical distributions to visually evaluate if the ratio is significantly different for EDA purposes [39].
- **The p-value.** This metric indicates the likelihood value of the observed data: if the value is below the threshold, null hypotheses are rejected [39]. We used a threshold value of zero.
- **Likelihood.** The likelihood parameter maps $L : \Theta \Longrightarrow \mathbf{R}$ or $\mathbf{R} : \Theta \Longrightarrow \mathbf{L}$ given by $\mathbf{R}[\Theta]|y] \Longrightarrow f_y[x]||[\Theta]$ or $\mathbf{L}[\Theta]||[x] \Longrightarrow f_y[y]||[\Theta]$. This metric computes the most probable value assigned to a specific feature using Θ as the hypothesis in $\mathbf{R}$ and $\mathbf{L}$ spaces. Inputs x compute the predictive values y using a predefined Θ parameter. With this, the likelihood represents the quantile probability (Prob (Q)) of correlated features used for forecasting.
- **Kurtosis.** This metric evaluates the probability of the predicted variables by describing the probability proportion. There are various techniques to compute the theoretical distribution of Kurtosis, and there are subjective manners of approximating it with relevant samples [39]. With Kurtosis results, higher values determine the presence of outliers. The Kurtosis is as follows:

$$Kurtosis[x] = [(\frac{x-\mu}{\sigma})^n], \tag{9}$$

where μ is the random selection of inputs x using a standard deviation σ following the constraints:

$$\sum_{i=1}\sum_{j=1}\frac{\mu^i}{\sigma^j}. \tag{10}$$

- **Jarque-Bera (JB) test.** This metric uses a Lagrange multiplier to test for data normality. The JB value tests if the distribution is normal by testing the Kurtosis to determine if features have a normal distribution.

A normal JB distribution will have symmetrical Kurtosis indicating the peaked in the distribution. We formulate the JB test as follows:

$$JB = n[\frac{\sqrt{b_1^2}}{6} + \frac{(b_2 - 3)^2}{24}], \tag{11}$$

Where: the sample size is n, $\sqrt{b_1}$ is the skewness sample, and b_2 is the Kurtosis coefficient.

- **Heteroscedasticity.** It checks the alternative hypothesis (H_A) versus the null hypothesis (H_0) [39]. With the alternative hypothesis, the empirical error is multiplying the function of various variables:

$$H_A : \sigma_1^2 = \sigma_2^2 * ... * \sigma_n^2. \tag{12}$$

However, a null hypothesis has equal error variances (homoscedasticity) [39]:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2. \tag{13}$$

- **Accuracy.** The balanced accuracy B_A of the ARIMA model is calculated with the following mathematical formulation [40]:

$$B_A = \frac{((TP/TP + FN) + (TN/(TN + FP)))}{2} \tag{14}$$

Where True Positive (TP) and True Negative (TN) denote correct classification, but misclassification is the False Positive (FP) and False Negative (FN) [39]. We used cross-validation rounds to build multiple training/testing subsets to decide which model is a suitable predictor of the growth in subscriber data (80% of the training set, 10% of the validation set, and 10% of the testing set).

3.4. Principal Component Analysis

ML models are used to address a range of prediction problems. The unsatisfactory prediction of ML classifiers originates from overfitting or underfitting features. The removal of irrelevant patterns guaranteed improved performance of the ARIMA computation. PCA was utilized on the UGRansome data to extract relevant patterns. PCA is a feature extraction methodology of this research. We denote PCA as follows:

$$\mathbf{P} = \frac{1}{t-1} + \sum_{t=1}^{k}([x(t)x(t)^T]), \tag{15}$$

with stochastic $x(t)$ and $t = 1, 2...l$ with n-dimensional inputs x having a probability matrix $\mathbf{P}$ of zero mean. The PCA formulation uses the covariance given in Equation 16 with a linear calculation of $x(t)$ inputs into $y(t)$ outputs:

$$y(t) = Q^T x(t), \tag{16}$$

Q is an orthogonal n*n matrix type where i represents the columns viewed as eigenvectors computed as follows:

$$y_i(t) = Q^T x(t), \tag{17}$$

The range in Equation 17 starts from 1 ... n where y_i is the new component of the i^{th} PCA. Table 2 depicts the PCA results using the UGRansome dataset.

3.5. ARIMA Predictor Model

Given a long period, $y_t[t = 1, 2, ..., T]$ of a spanning time series traffic, the aim is to come up with a new scheme that works well for predicting the future outcomes H. We define $S = [1, 2, ..., T]$ as the sequence of timestamp with time series y_t.

Table 2. The PCA results using the UGRansome

Attack	Feature	Total
Blacklist	Timestamp	2,761
Spam	IP address	7,425
Scan	Flag	1,559
SSH	Prediction	7,293
Botnet	Threats	4,765
Total	-	**23,803**

The prediction of the problem can be written as $f[\Theta, \sum |y_t, t \in S]$, where the parameter is f, the global parameters Θ, and the covariance matrix $\sum$. However, the time series data is divided into different sub-series (k) having contiguous time intervals (Equation 18):

$$S = \sum_{k=1}^{K} S_k, \tag{18}$$

where S_k extracts the kTh sub-series timestamps and we posit $T = \sum_{k=1}^{K} T_k$. With this assumption, the predictor estimator of the sub-problem is shown in Equation 19:

$$f[\Theta, \sum |y_t, t \in S] = g[f_1, \sum_{1} |y_t, t \in S_1] ... [f_k, \sum_{k} |y_t, t \in S_k], \tag{19}$$

where f_k represents the estimator function for the kTh sub-series and g the combination. The estimation was merged before the prediction. The idea is to use $g(.)$ as a single mean parameter, and our computational framework could be viewed as an averaging ML algorithm (Figure 9).

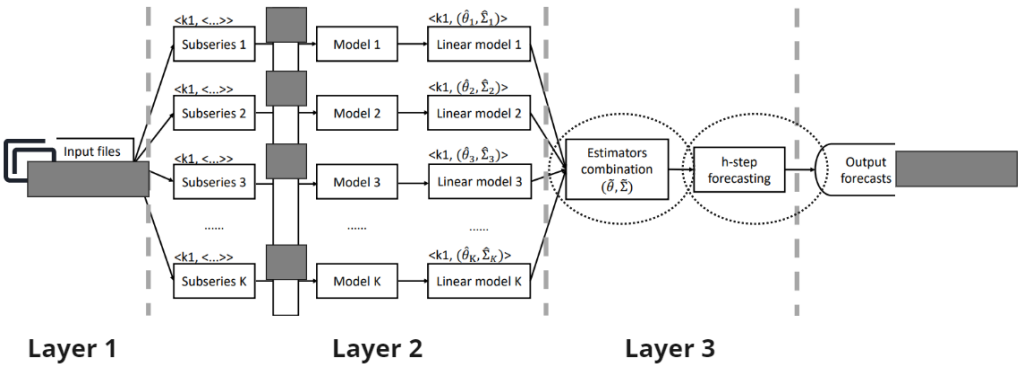


Figure 9. The ARIMA model.

Figure 9 outlines the proposed ARIMA model to forecast the growth in subscriber data. The timestamps of historical data were recorded in the IDS before being processed by the ARIMA model. In simple terms, the proposed ARIMA model consists of the following phases:

- **Phase1: Preprocessing.** Subdivide the time series data into the K sub-series.
- **Phase 2: Modeling.** Train the algorithm using sub-series data by assuming that the IDS of the sub-series remains constant.
- **Phase 3: Linear transformation.** Translate the trained algorithm in phase 2 into linear representations k.
- **Phase 4: Estimator combination.** The obtained local estimator from phase 3 minimizes global losses parameters described in Section 2.3.
- **Phase 5: Prediction.** Predict the next observations H by utilizing the merged estimator's parameters presented in Equation 18 and 19.

We used available hourly-based timestamps to create a new set of timestamps (ts) used in the ARIMA prediction. The following formulation was used to predict new timestamps (Equation 20):

$$Predicted_{ts} = LastEpochTimestamps + n * 3,600 \quad (20)$$

where $n = range(1, 48)$. This computation provides new predicted timestamps on an hourly basis for the next 48 hours.

3.6. Computational Environment

The IDS used to build the subscriber data is installed on a DBeaver database. DBeaver is a database monitoring software that manipulates telecommunication data like Deep Packet Inspection (DPI). It can be used to build analytical dashboards from various data storage. In this article, we exported the DBeaver data in an appropriate CSV format (Figure 4). The ARIMA model is computed with Python on Jupyter, and the invoked ML packages used to implement EDA are shown in Algorithm 1. In Figure 10 ARIMA is presented and the employed computing environment is illustrated in Table 3.

Algorithm 1 The EDA algorithm

Require: pip install -u dataprep
Ensure: from datetime import datetime
Ensure: import numpy as np
Ensure: import pandas as pd
Ensure: import matplotlib.pylab as plt
Ensure: % matplotlib inline
Ensure: from matplotlib.pylab import rcParams
Ensure: from statsmodels.tsa.stattools import adfuller
Ensure: !pip install pmdarima −− quiet
Ensure: import pmdarima as pm
Ensure: from dataprep.datasets import load dataset
Ensure: from dataprep.eda
Ensure: import create report
Require: read dataset()
Require: create report()
Require: show browser()

```
def forecast(ARIMA_model, periods=24):
    # Forecast
    n_periods = periods
    fitted, confint = ARIMA_model.predict(n_periods=n_periods, return_conf_int=True)
    index_of_fc = pd.date_range(df.index[-1], periods = n_periods, freq='H')
     # make series for plotting purpose
    fitted_series = pd.Series(fitted, index=index_of_fc)
    lower_series = pd.Series(confint[:, 0], index=index_of_fc)
    upper_series = pd.Series(confint[:, 1], index=index_of_fc)

    # Plot
    plt.figure(figsize=(15,7))
    #plt.plot(df[col], color='#1f76b4')
    plt.plot(fitted_series, color='darkgreen')
    #plt.fill_between(lower_series.index,
                     #lower_series,
                     #upper_series,
                     #color='k', alpha=.15)

    plt.title("ARIMA/SARIMA - Forecast of Tpt_out")
    plt.show()

forecast(ARIMA_model)
ARIMA_model.summary()
```

Figure 10. The forecasting function.

Table 3. Framework specification

Node	Specification
RAM	39 GB
Service	Jupyter & DBeaver
ML algorithm	ARIMA & CNN
System	64-bits
Processor	2.60 GHz
Dataset	Subscriber data & UGRansome
Operating system	Windows & Linux
CPU	Intel i7-10
Language	Python

3.7. Feature Extraction

There are different reasons causing duplication in a dataset, among which are imperfections in the data collection process and the properties of features, but feature extraction solved redundancy dimensions. Features projected into a new space have lower dimensionality. Examples of such techniques include Linear Discriminant Analysis (LDA), Canonical Correlation Analysis (CCA), and Principal Component Analysis [34]. We have used PCA to improve the predictive performance of the ARIMA model. The PCA lowered computational complexity, built generalizable models, and optimized the storage space. To address the redundancy issue, the PCA selected a subset of relevant patterns from the original dataset based on their relevance. We present the PCA results in Table 4 where the final dataset with the description of each attribute is presented.

Table 4. Extracted features

Number	Attribute	Description	Type
1	Timestamp	Traffic duration	Numeric
2	Protocol	Communication rule	Categorical
3	Flag	Network state	Categorical
4	IP address	Unique address	Categorical
5	Network traffic	Periodic network flow	Numeric
6	Threat	Novel malware	Categorical
7	Port	Communication port	Numeric
8	Expended address	Malware address	Categorical
9	Seed address	Malware address	Categorical
10	Cluster	Group assigned	Numeric
11	Ransomware	Novel malware	Categorical
12	Prediction	Novel malware class	Categorical

The prediction attribute facilitates the forecasting of any ML model to predict the category of novel intrusion. Our final dataset has 12 variables with 180,564 observations (Table 4). Consequently, if the deviation degree of a variable is high or low enough, it is considered an abnormality. The feature extraction using UGRansome leads to improved performance, higher prediction accuracy, minimized computational time, and efficient model interpretability.

However, we did not apply feature extraction on the subscriber data because it has no redundant features.

4. Results

This section compares the ARIMA model performance using the subscriber and the UGRansome datasets. Figure 11 shows the format of the subscriber data following a linear distribution compared to the UGRansome timestamps.

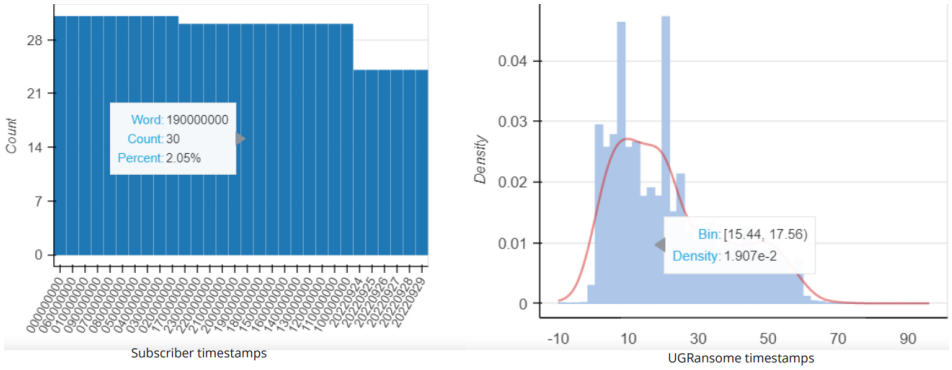


Figure 11. The timestamp density comparison.

Figure 12 portrays a distribution of incoming and outgoing throughput of the subscriber data compared to the UGRansome port traffic (5066-5068). Each attack flow is also depicted. The figure depicts NerisBonet threats with fewer traffic.

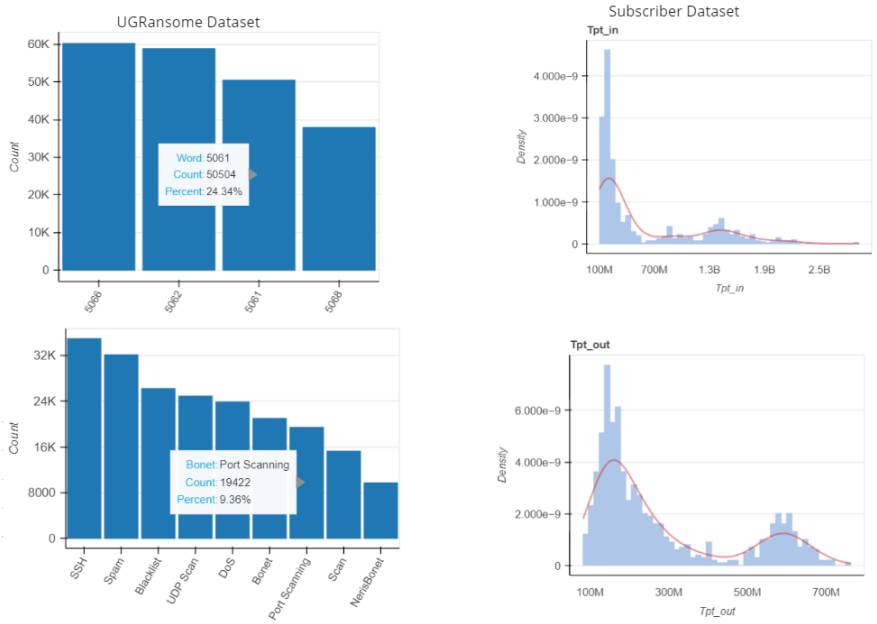


Figure 12. Additional features comparison.

This result reveals a time series forecasting property of both used datasets. However, the UGRansome has more distributed or dependent variables (Figure 13). The correlation of throughput is in Figure 14. This plot indicates the linear distribution of predicted values. The summary of the ARIMA model using the subscriber data is presented in Figure 15. The summary confirms that the prediction of subscriber data will have an increased mean or standard deviation given the likelihood, probability (Prob(Q)), and Kurtosis values. In the next section, we will compare the subscriber data with the UGRansome using the Dickey-Fuller Test (DFT) results.

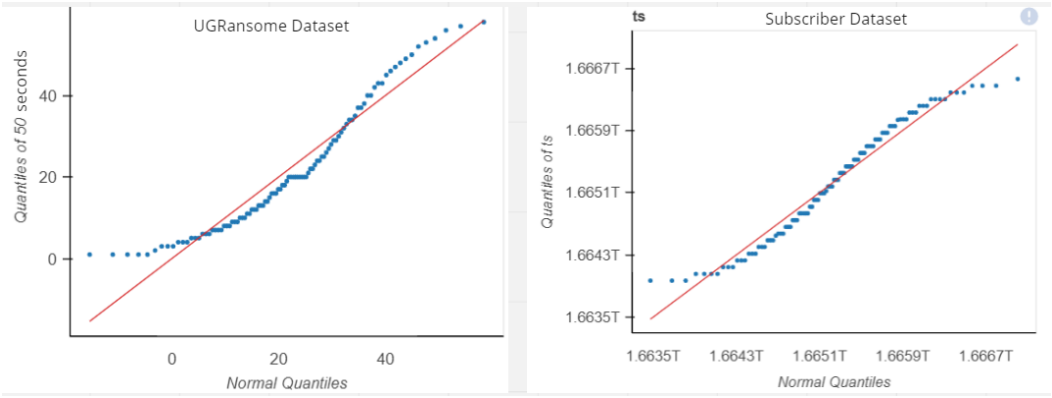


Figure 13. The normal Q-Q results.

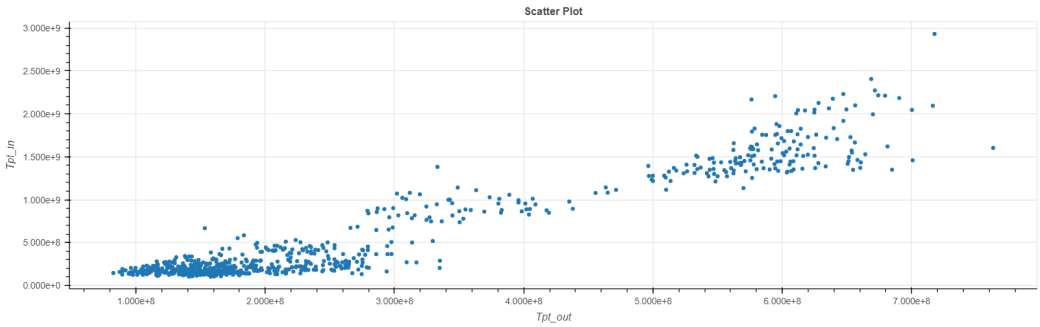


Figure 14. The throughput correlation.

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	732
Model:	SARIMAX(3, 0, 1)	**Log Likelihood**	-14399.470
Date:	Mon, 31 Oct 2022	**AIC**	28810.941
Time:	17:51:26	**BIC**	28838.515
Sample:	0	**HQIC**	28821.578
	- 732		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	4.899e+07	5.38e-09	9.1e+15	0.000	4.9e+07	4.9e+07
ar.L1	1.1487	0.103	11.154	0.000	0.947	1.351
ar.L2	-0.1228	0.135	-0.912	0.362	-0.387	0.141
ar.L3	-0.1855	0.051	-3.621	0.000	-0.286	-0.085
ma.L1	-0.1868	0.118	-1.588	0.112	-0.417	0.044
sigma2	7.181e+15	3.07e-17	2.34e+32	0.000	7.18e+15	7.18e+15

Ljung-Box (L1) (Q):	0.02	**Jarque-Bera (JB):**	376.22
Prob(Q):	0.88	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.35	**Skew:**	1.22
Prob(H) (two-sided):	0.02	**Kurtosis:**	5.54

Figure 15. The SARIMA model summary.

4.1. Dickey Fuller Test

The DFT results are in Table 5. A p-value of 0.007 with an accuracy of 90% was obtained. This result supports our hypothesis predicting an increase in subscriber data growth. As such, (i) the UGRansome used more iterations due to its size surpassing the subscriber dataset, (ii) the balanced accuracy of the DFT reached 81% of accuracy, and (iii) the dataset size has not to effect on the prediction performance. The residual and correlogram are in Figure 16 with the seasonality of the throughput.

Both experimental datasets achieved the same predictive accuracy value of 90%. However, the data usage growth prediction is illustrated in Figure 17 using the ARIMA model that predicts a growth of 3 Mbps at a specific timestamp. UNIX timestamps of the subscriber data predicted the maximum data growth using the ARIMA model. In Figure 18, ARIMA predicts a maximal subscriber data usage growth of 14 Gbps (where blue denotes actual data, orange is the predicted ARIMA data, and green is the future predicted values). The predicted mean and standard deviation are in Figure 19. The original data represents predicted values, but ARIMA approximated a mean and standard deviation from these values. ARIMA predicted a maximum mean value of five Mbps with a standard deviation of two Mbps. The results show mean and standard deviation values lower than the original predicted values.

Table 5. The DFT results

Dataset	**Test statistic**	**p-value**	**Iteration**	**Accuracy**
Subscriber data	-3.537,879	0.007,066	20	90.567%
UGRansome data	-9.876,982	0.0,008,044	342	90.456%
	Correlogram	**ADF**	**Q-Q**	
Subscriber training set	0.9	0.8	0.9	90.398%
UGRansome training set	0.8	0.9	0.7	89.453%
Subscriber testing set	0.8	0.9	0.9	91.348%
UGRansome testing set	0.8	0.8	0.9	88.298%
	Features Total	**Mean**	**Deviation**	
Subscriber data	700	54.23	22.45	92.351%
UGRansome data	8,932	75.32	46.3	88.527%
Subscriber testing set	400	12.6	6.7	94%
UGRansome testing set	4,765	26.87	39.65	88%
Balanced Accuracy	-	-	-	**81%**
Balanced Features	**3,699**	-	-	-
Balanced Mean	-	**41.75**	-	-
Balanced Deviation	-	-	**28.25**	-

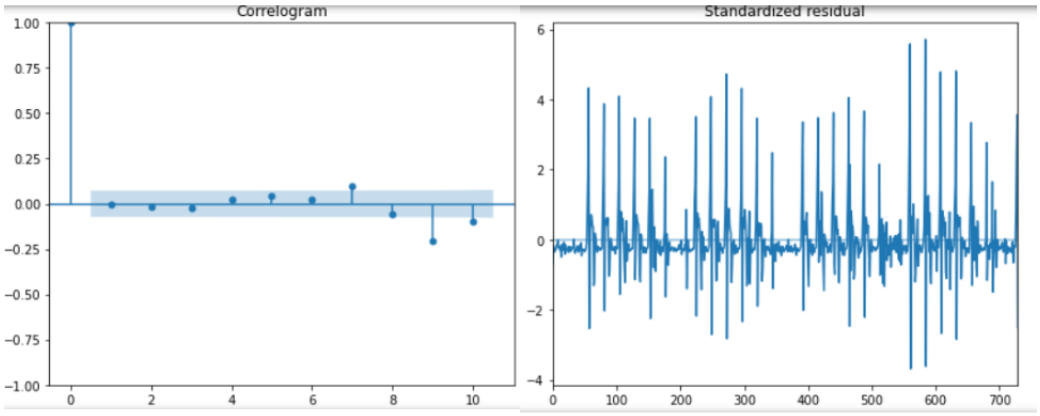


Figure 16. The standardized residual and correlogram.

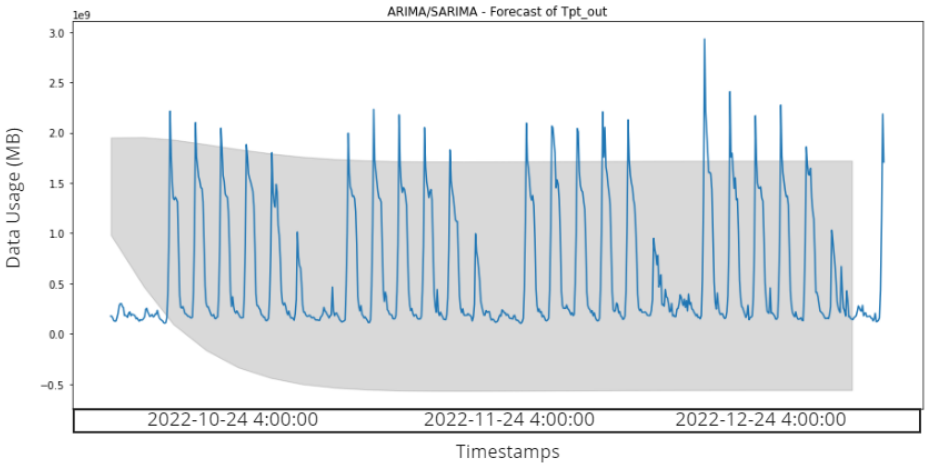


Figure 17. The ARIMA prediction.

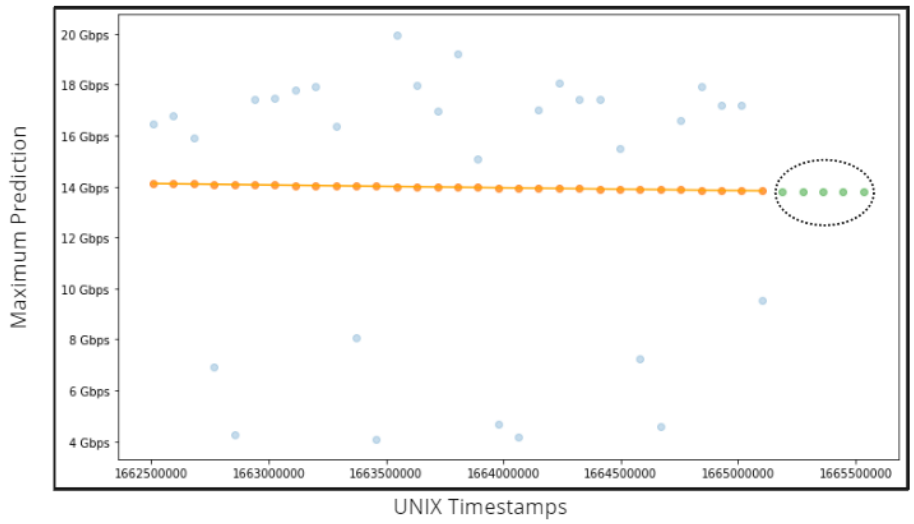


Figure 18. The maximum data usage prediction.

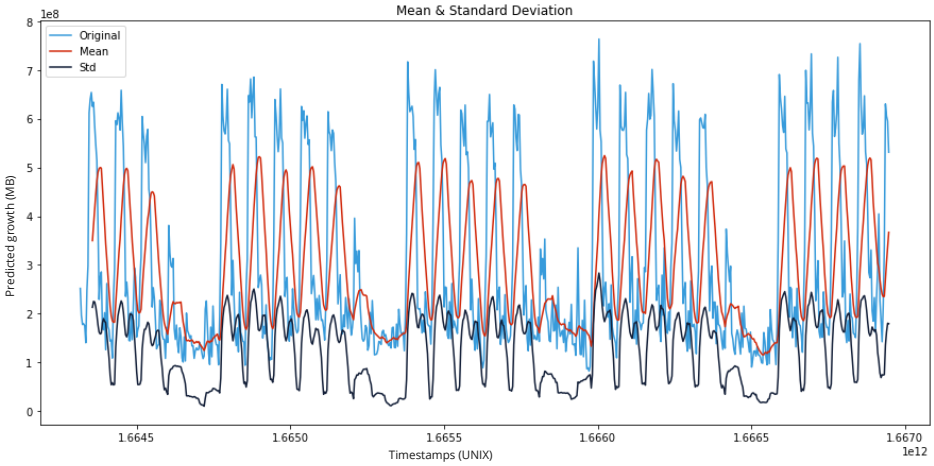


Figure 19. The prediction of the mean and standard deviation.

4.2. The Comparative Results of ARIMA and CNN

The Convolutional Neural Network (CNN) is compared to ARIMA using the subscriber and UGRansome datasets. The CNN depends on the predicted timestamps. In what follows, we present the CNN results compared to the results obtained by the ARIMA. The prediction includes 30 to 60 days. Considering a single pattern of features as depicted in Figure 20, the CNN weights features to enable the learning trend of particular observations. We have various observations, so we merge their outputs into a connected layer. Our CNN architecture uses binary convolutions (with 70 and 30 filters) and a densely connected layer of 130 neurons with the activation function (RELU) (see Figure 20). This unit has six connected layers representing auxiliary outputs (Figure 20). Each layer predicts and passes the prediction value to the next layer which predicts growth in the subscriber data usage until the final layer produces long-term forecasting. We use each layer to predict in advance some additional days. We used grid search to detect the optimal number of filters, convolutions, connected layers, and drop-out rate. For each layer $k \in [1, 2, 3, 4, 5, 6]$ we added a Mean-Squared Error (MSE) and loss function (Figure 20). Each layer k aims to produce future forecasting for more than 14 days ahead:

$$k_{loss} \Longrightarrow \frac{1}{n}\sum_{i}(y_i, 14_k - tanh(\hat{y}_i, 14_k))^2 \tag{21}$$

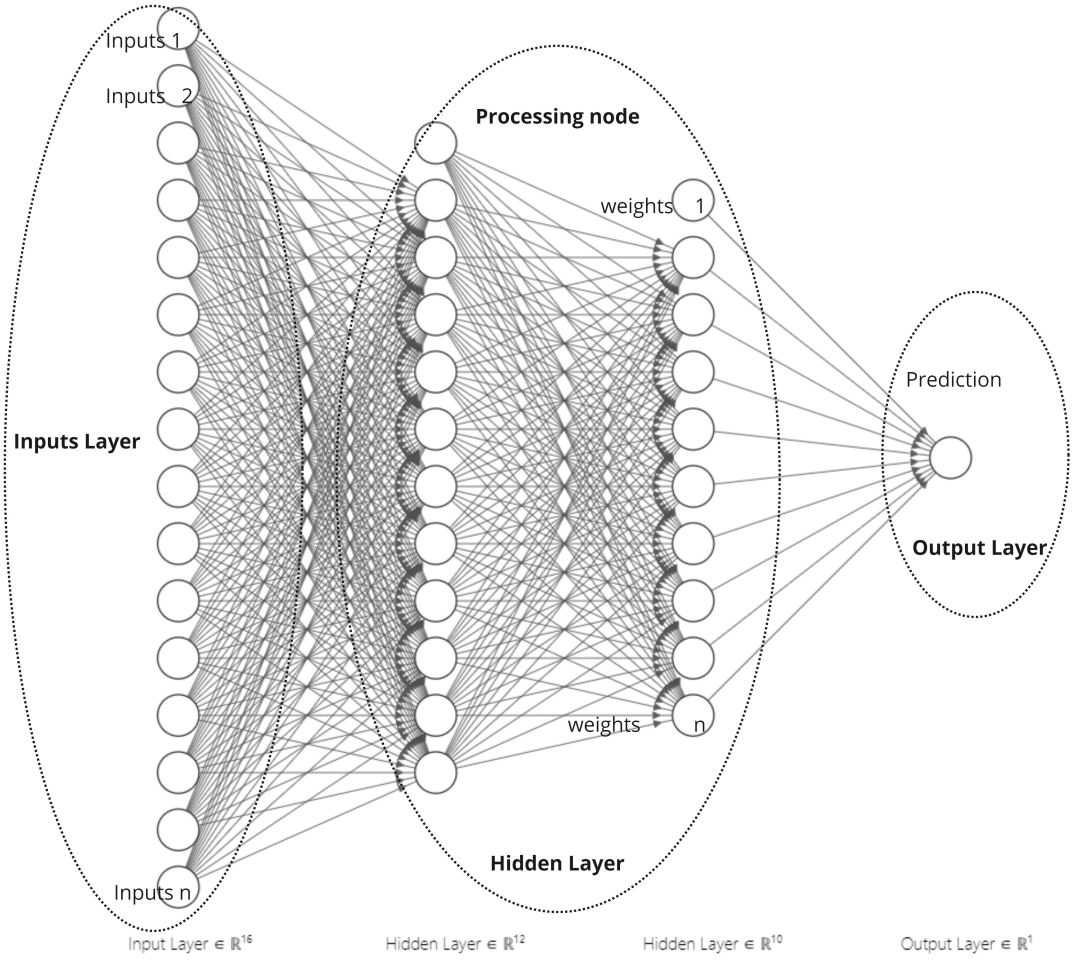


Figure 20. The CNN architecture.

where $y_i, 14_k$ represents the features of time-series i, $\hat{y}_i, k$ the kTh layer's forecasting the value, and tanh the function restricting the range of $\hat{y}_i, k$ to $[-1, 1]$.

With this, we can reformulate Equation 21 to the weighted sum and minimize the loss by decreasing λ_k values:

$$Min_{loss} \Longrightarrow \frac{1}{k}\sum_{k=1}^{K}(\frac{k_{loss}}{\lambda_{loss}}) \tag{22}$$

The comparative results of the ARIMA and CNN models are in Table 6.

Table 6. The CNN and ARIMA results

Dataset	Features	p-value	CNN	ARIMA
Subscriber data	450	0.006,055	85.8%	**92.67%**
UGRansome data	120,000	0.0,006,043	88.9%	**91.65%**
Subscriber testing set	300	0.008	86.3%	**94.8%**
UGRansome testing set	60,500	0.007	88%	**95.3%**
Balance	45,312	0,005	87,25%	**93.605%**

The CNN is compared to ARIMA using experimental datasets. We used four samples. The first sample was the subscriber dataset, where the ARIMA model obtained 92% of accuracy and outperformed the CNN. The second sample was the UGRansome dataset containing more features, but the ARIMA model surpassed the CNN with 91% of accuracy. The third sample was the testing sample of the subscriber data where the ARIMA achieved 94%, and in the last sample, the ARIMA accuracy outperformed the CNN with 95% of accuracy. Overall, the ARIMA model achieved the best results in all undertaken comparisons. The ARIMA model performed better with the UGRansome data, and this was due to the nature of seasonal network traffic. We computed our models on fewer features of the subscriber data without producing poor results. We believe this is due to time series data properties which improve the balanced accuracy with 93% of accuracy.

5. Conclusion and Discussion

Insights retrieval from subscriber data impacts the telecommunication landscape to facilitate information management and assist decision-makers in predicting the future using ML techniques. We explore time series forecasting analysis and predict subscriber usage trends on the network using the ARIMA model. The unknown forecasting value used by ARIMA relied on historical data. However, we used the data storage to build the subscriber dataset using hourly traffic statistics. We used various metrics to evaluate the ARIMA model. For instance, the normal Q-Q, standardized residual, theoretical quantile, correlogram, and accuracy. UGRansome was used to compare the obtained results that demonstrate similar accuracy values of 90% using the ARIMA model. The subscriber data was not stationary but linear and seasonal. In the experimentation, ARIMA was compared to the Convolutional Neural Network (CNN) and achieved the best results with the UGRansome data. We have used an NSDM environment with subscriber data in a secure environment to retrieve relevant patterns such as timestamps and incoming/outgoing throughout to build the subscriber dataset. The variation of the auto-regressive and moving average components identified the most optimal features for obtaining precise predictive values. In addition, the subscriber data have normal distributions, but the UGRansome has more dependent variables. The ARIMA model predicted a growth of 3 Mbps with a maximum data usage growth of 14 Gbps. Future works can concentrate on seasonality to understand changes in subscriber data or merge the UGRansome with the subscriber data to build a compacted dataset. This dataset will be used with Deep Learning to forecast network attacks in real-time. One shortcoming of our framework is the modification of various forward forecasting periods. However, Deep Learning models can address this issue. In the future, we will optimize the CNN to change the forward forecasting period without retraining the model.

Author Contributions: Conceptualization, M.N., and J.P.v.D; writing—original draft preparation, M.N; supervision, J.P.v.D; writing—review and editing, J.P.v.D; experiments and testing, M.N. All authors have read and agreed to the published version of the manuscript.

Funding: The authors wish to extend their sincere appreciation and gratitude to the Editor-in-Chief, Prof. Dr. Antonio Gil Bravo (Public University of Navarra, Spain), for the invitation to contribute with this article entitled to a fee waiver discount.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset and code used can be obtained upon request or downloaded at https://www.researchgate.net/publication/342200905_An_Ensemble_Learning_Framework_for_Anomaly_Detection_in_the_Existing_Network_Intrusion_Detection_Landscape (Public Files/Ugransome.zip and subscriber data.csv). The code is under (Public Files/ARIMA). Accessed on 2022-11-01.

Conflicts of Interest: The authors declare no conflict of interests.

References

1. Theodoridis, G.; Tsadiras, A. Applying machine learning techniques to predict and explain subscriber churn of an online drug information platform. *Neural Computing and Applications* **2022**, pp. 1–14. https://doi.org/10.1007/s00521-022-07603-9.
2. Ghaderi, A.; Movahedi, Z. Joint Latency and Energy-aware Data Management Layer for Industrial IoT. In Proceedings of the 2022 8th International Conference on Web Research (ICWR). IEEE, 2022, pp. 70–75. https://doi.org/10.1109/ICWR54782.2022.9786229.
3. Jin, X.B.; Gong, W.T.; Kong, J.L.; Bai, Y.T.; Su, T.L. A variational Bayesian deep network with data self-screening layer for massive time-series data forecasting. *Entropy* **2022**, *24*, 335. https://doi.org/https://doi.org/10.3390/e24030335.
4. Li, X.; Petropoulos, F.; Kang, Y. Improving forecasting by subsampling seasonal time series. *International Journal of Production Research* **2022**, pp. 1–17. https://doi.org/https://doi.org/10.1080/00207543.2021.2022800.
5. Kumar, R.; Kumar, P.; Kumar, Y. Multi-step time series analysis and forecasting strategy using ARIMA and evolutionary algorithms. *International Journal of Information Technology* **2022**, *14*, 359–373. https://doi.org/https://doi.org/10.1007/s41870-021-00741-8.
6. Tan, C.W.; Bergmeir, C.; Petitjean, F.; Webb, G.I. Time series extrinsic regression. *arXiv preprint arXiv:2006.12672* **2020**. https://doi.org/https://doi.org/10.1007/s10618-021-00745-9.
7. Goldsmith, J.; Scheipl, F. Estimator selection and combination in scalar-on-function regression. *Computational Statistics & Data Analysis* **2014**, *70*, 362–372. https://doi.org/https://doi.org/10.1016/j.csda.2013.10.009.
8. Pimentel, M.A.; Charlton, P.H.; Clifton, D.A. Probabilistic estimation of respiratory rate from wearable sensors. In *Wearable electronics sensors*; Springer, 2015; pp. 241–262. https://doi.org/https://doi.org/10.1007/978-3-319-18191-2_10.
9. Zheng, Y.; Liu, Q.; Chen, E.; Ge, Y.; Zhao, J.L. Time series classification using multi-channels deep convolutional neural networks. In Proceedings of the International conference on web-age information management. Springer, 2014, pp. 298–310.
10. Yang, J.; Nguyen, M.N.; San, P.P.; Li, X.L.; Krishnaswamy, S. Deep convolutional neural networks on multichannel time series for human activity recognition. In Proceedings of the Twenty-fourth international joint conference on artificial intelligence, 2015.
11. Okita, T.; Inoue, S. Recognition of multiple overlapping activities using compositional CNN-LSTM model. In Proceedings of the Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, 2017, pp. 165–168. https://doi.org/https://doi.org/10.1145/3123024.3123095.
12. Wang, J.; Long, Q.; Liu, K.; Xie, Y.; et al. Human action recognition on cellphone using compositional bidir-lstm-cnn networks. In Proceedings of the 2019 International Conference on Computer, Network, Communication and Information Systems (CNCI 2019). Atlantis Press, 2019, pp. 687–692.

13. Snow, D. AtsPy: Automated Time Series Forecasting in Python. *Available at SSRN 3580631* **2020**. https://doi.org/http://dx.doi.org/10.2139/ssrn.3580631.
14. Mode, G.R.; Hoque, K.A. Adversarial examples in deep learning for multivariate time series regression. In Proceedings of the 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). IEEE, 2020, pp. 1–10. https://doi.org/10.1109/AIPR50011.2020.9425190.
15. Antsfeld, L.; Chidlovskii, B.; Borisov, D. Magnetic sensor based indoor positioning by multi-channel deep regression. In Proceedings of the Proceedings of the 18th Conference on Embedded Networked Sensor Systems, 2020, pp. 707–708. https://doi.org/https://doi.org/10.1145/3384419.3430419.
16. Mehtab, S.; Sen, J.; Dasgupta, S. Robust analysis of stock price time series using CNN and LSTM-based deep learning models. In Proceedings of the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020, pp. 1481–1486. https://doi.org/10.1109/ICECA49313.2020.9297652.
17. Mirko, K.; Kantelhardt, J.W. Hadoop. TS: large-scale time-series processing. *International Journal of Computer Applications* **2013**, *74*. https://doi.org/https://doi.org/10.1007/s41870-021-00741-8.
18. Li, L.; Noorian, F.; Moss, D.J.; Leong, P.H. Rolling window time series prediction using MapReduce. In Proceedings of the Proceedings of the 2014 IEEE 15th international conference on information reuse and integration (IEEE IRI 2014). IEEE, 2014, pp. 757–764. https://doi.org/10.1109/IRI.2014.7051965.
19. Talavera-Llames, R.; Pérez-Chacón, R.; Troncoso, A.; Martínez-Álvarez, F. Big data time series forecasting based on nearest neighbours distributed computing with Spark. *Knowledge-Based Systems* **2018**, *161*, 12–25. https://doi.org/https://doi.org/10.1016/j.knosys.2018.07.026.
20. Galicia, A.; Torres, J.F.; Martínez-Álvarez, F.; Troncoso, A. A novel Spark-based multi-step forecasting algorithm for big data time series. *Information Sciences* **2018**, *467*, 800–818. https://doi.org/https://doi.org/10.1016/j.ins.2018.06.010.
21. Petropoulos, F.; Apiletti, D.; Assimakopoulos, V.; Babai, M.Z.; Barrow, D.K.; Taieb, S.B.; Bergmeir, C.; Bessa, R.J.; Bijak, J.; Boylan, J.E.; et al. Forecasting: theory and practice. *International Journal of Forecasting* **2022**.
22. Shamir, O.; Srebro, N.; Zhang, T. Communication-efficient distributed optimization using an approximate newton-type method. In Proceedings of the International conference on machine learning. PMLR, 2014, pp. 1000–1008.
23. Wang, J.; Kolar, M.; Srebro, N.; Zhang, T. Efficient distributed learning with sparsity. In Proceedings of the International conference on machine learning. PMLR, 2017, pp. 3636–3645.
24. Jordan, M.I.; Lee, J.D.; Yang, Y. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* **2018**. https://doi.org/https://doi.org/10.1080/01621459.2018.1429274.
25. Chen, X.; Liu, W.; Zhang, Y. Quantile regression under memory constraint. *The Annals of Statistics* **2019**, *47*, 3244–3273. https://doi.org/10.1214/18-AOS1777.
26. Ryu, E.K.; Yin, W. *Large-Scale Convex Optimization*; Cambridge University Press, 2022.
27. Challu, C.; Olivares, K.G.; Oreshkin, B.N.; Garza, F.; Mergenthaler, M.; Dubrawski, A. N-hits: Neural hierarchical interpolation for time series forecasting. *arXiv preprint arXiv:2201.12886* **2022**. https://doi.org/https://doi.org/10.48550/arXiv.2201.12886.
28. Fernández, J.D.; Menci, S.P.; Lee, C.M.; Rieger, A.; Fridgen, G. Privacy-preserving federated learning for residential short-term load forecasting. *Applied Energy* **2022**, *326*, 119915. https://doi.org/https://doi.org/10.1016/j.apenergy.2022.119915.
29. Bennett, S.; Clarkson, J. Time series prediction under distribution shift using differentiable forgetting. *arXiv preprint arXiv:2207.11486* **2022**. https://doi.org/https://doi.org/10.48550/arXiv.2207.11486.
30. Mehdi, H.; Pooranian, Z.; Vinueza Naranjo, P.G. Cloud traffic prediction based on fuzzy ARIMA model with low dependence on historical data. *Transactions on Emerging Telecommunications Technologies* **2022**, *33*, e3731. https://doi.org/https://doi.org/10.1002/ett.3731.
31. Xiao, R.; Feng, Y.; Yan, L.; Ma, Y. Predict stock prices with ARIMA and LSTM. *arXiv preprint arXiv:2209.02407* **2022**.
32. Wang, X.; Kang, Y.; Hyndman, R.J.; Li, F. Distributed ARIMA models for ultra-long time series. *International Journal of Forecasting* **2022**. https://doi.org/https://doi.org/10.1016/j.ijforecast.2022.05.001.
33. Nkongolo, M.; van Deventer, J.P.; Kasongo, S.M. The Application of Cyclostationary Malware Detection Using Boruta and PCA. In Proceedings of the Computer Networks and Inventive Communication Technologies; Smys, S.; Lafata, P.; Palanisamy, R.; Kamel, K.A., Eds.; Springer

Nature Singapore: Singapore, 2023; pp. 547–562. https://doi.org/https://doi.org/10.1007/978-981-19-3035-5_41.

34. Nkongolo, M.; Van Deventer, J.P.; Kasongo, S.M.; Zahra, S.R.; Kipongo, J. A Cloud Based Optimization Method for Zero-Day Threats Detection Using Genetic Algorithm and Ensemble Learning. *Electronics* **2022**, *11*. https://doi.org/10.3390/electronics11111749.
35. Nkongolo, M.; van Deventer, J.P.; Kasongo, S.M. UGRansome1819: A Novel Dataset for Anomaly Detection and Zero-Day Threats. *Information* **2021**, *12*. https://doi.org/10.3390/info12100405.
36. Chao, H.L.; Liao, W. Fair scheduling in mobile ad hoc networks with channel errors. *IEEE transactions on wireless communications* **2005**, *4*, 1254–1263. https://doi.org/10.1109/TWC.2004.842942.
37. Suthar, F.; Patel, N.; Khanna, S. A Signature-Based Botnet (Emotet) Detection Mechanism. *International Journal of Engineering Trends and Technology* **2022**, pp. 185–193.
38. Kotu, V.; Deshpande, B. Chapter 3 - Data Exploration. In *Data Science (Second Edition)*, Second Edition ed.; Kotu, V.; Deshpande, B., Eds.; Morgan Kaufmann, 2019; pp. 39–64. https://doi.org/https://doi.org/10.1016/B978-0-12-814761-0.00003-4.
39. Ij, H. Statistics versus machine learning. *Nat Methods* **2018**, *15*, 233.
40. Nkongolo, M. Classifying search results using neural networks and anomaly detection. *Educor Multidisciplinary Journal* **2018**, *2*, 102–127. https://doi.org/https://hdl.handle.net/10520/EJC-13d317b93a.

eng

Submit to *Eng*

Review for *Eng*

Journal Menu

- *Eng* Home
- Aims & Scope
- Editorial Board
- Reviewer Board
- Topical Advisory Panel
- Instructions for Authors
- Special Issues
- Topics
- Sections
- Article Processing Charge
- Indexing & Archiving
- Most Cited & Viewed
- Journal History
- Editorial Office

Journal Browser

volume

issue

Go

Forthcoming issue

Current issue

Vol. 3 (2022) Vol. 1 (2020)

Vol. 2 (2021)

Editorial Board

- Chemical, Civil and Environmental Engineering Section
- Materials Engineering Section
- Electrical and Electronic Engineering Section

Members (86)

Search by first name, last name, affiliation, interest...

Prof. Dr. Antonio Gil Bravo Website SciProfiles

Editor-in-Chief

Science Department, Public University of Navarra, Building Los Acebos, Campus of Arrosadia, E-31006 Pamplona, Spain

Interests: preparation, characterization and catalytic performance of metal supported nanocatalysts; catalytic combustion of volatile organic compounds (VOC); porous and surface properties of solids; gas adsorption; energy storage; pollutant adsorption; environmental management; industrial waste valorization; circular economy

Special Issues, Collections and Topics in MDPI journals

Prof. Dr. Leszek Adam Dobrzanski Website SciProfiles

Section Editor-in-Chief

Medical and Dental Engineering Centre for Research, Design and Production ASKLEPIOS, 44-100 Gliwice, Poland

Interests: materials engineering; nanotechnology; biomaterials; medical; dental; manufacturing and surface engineering; machine building and automation; management and organization

Special Issues, Collections and Topics in MDPI journals

Dr. George Z. Papageorgiou Website SciProfiles

Section Editor-in-Chief

Department of Chemistry, University of Ioannina, 45110 Ioannina, Greece

Interests: sustainable polymers; biobased materials and chemicals; thermal processes; thermal analysis; polymer engineering; biodegradation; green engineering; bioresources and biorefinery; polymer wastes; recycling

Special Issues, Collections and Topics in MDPI journals

Dr. Amor Abdelkader Website SciProfiles

Editorial Board Member

1. Department of Materials Science and Metallurgy, University of Cambridge, Cambridge CB3 0FS, UK
2. Department of Design and Engineering, Faculty of Science & Technology, Bournemouth University, Poole BH12 5BB, UK

Interests: production, processing and applications of nanomaterials; 2D materials technology; electrochemical energy storage devices

Special Issues, Collections and Topics in MDPI journals

Dr. Paolo Addesso Website SciProfiles

Editorial Board Member

DIEM, University of Salerno, 84084 Fisciano, Italy

Interests: remote sensing; image processing; signal processing; sequential Bayesian estimation; estimation theory; detection theory; statistical signal processing; fractal models; data fusion; gravitational waves; localization; nonlinear devices; sensor networks

Special Issues, Collections and Topics in MDPI journals

Dr. Rafik Addou Website SciProfiles

Editorial Board Member

School of Chemical, Biological and Environmental Engineering, Oregon State University, Corvallis, OR 97331, USA

Interests: 2D materials; scanning probe microscopy; XPS, nanofabrication; thin films, interface and surface science

Special Issues, Collections and Topics in MDPI journals

Dr. Shatirah Akib Website SciProfiles

Editorial Board Member

Department of Civil Engineering, School of Architecture, Design and the Built Environment, Nottingham Trent University, Nottingham NG1 4FQ, UK

Interests: fluid mechanics; hydraulic structure engineering; hydraulic and water engineering; river and coastal engineering; renewable energy; natural disaster

Special Issues, Collections and Topics in MDPI journals

Dr. Muthanna H. Al-Dahhan Website SciProfiles

Editorial Board Member

Department of Chemical and Biochemical Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA

Interests: multiphase reaction engineering; advanced measurement techniques; clean and alternative energy and environment

Special Issues, Collections and Topics in MDPI journals

Prof. Dr. Abdeltif Amrane Website SciProfiles

Editorial Board Member

Rennes Institute of Chemical Sciences, University of Rennes, CEDEX 7, 35708 Rennes, France