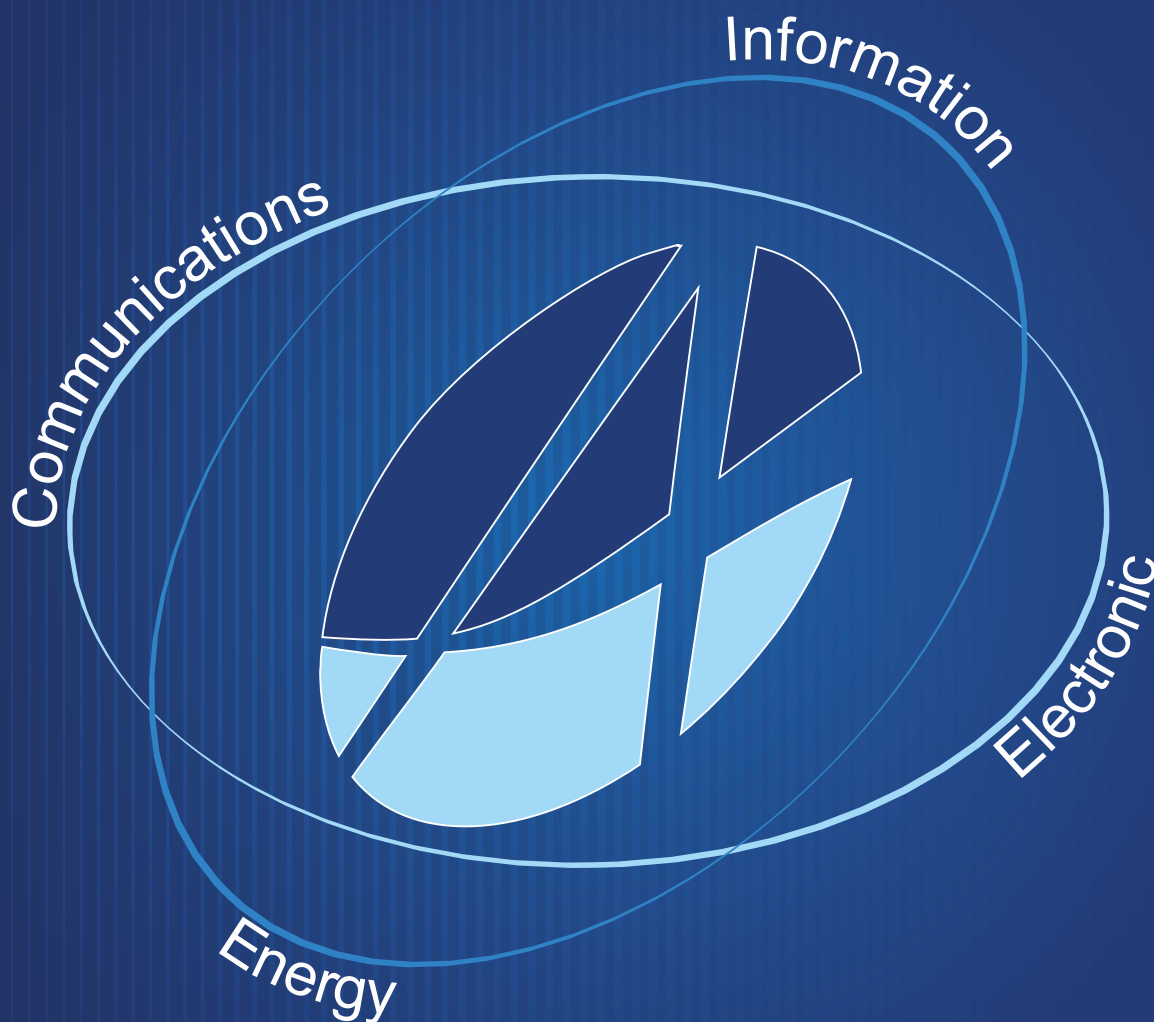


The Journal of **CIEES**

Communications, Information, Electronic & Energy Systems

Print ISSN: 2738-7283
Online ISSN: 2738-7291

Year 2026
Volume 6 (1)



No liability is assumed by the editorial staff for the content of scientific papers and opinions published in this volume. They represent the author's point of view.

Editorial

As a result of widening the spectrum of communication, dissemination and implementation of scientific research results, in 2020 was founded the association journal (magazine), issued under the name The Journal of CIEES appearing both in print (ISSN: 2738-7283), and online (ISSN: 2738-7291) format with DOI: <https://doi.org/10.48149/jciees>.

The Journal of CIEES is published in English and it is open access, peer-reviewed international scientific journal publishing papers submitted by authors themselves for publishing and invited by the Editorial Board. The Journal of CIEES covers all the essential branches of theories, technologies, systems, tools, applications, work in progress, and experiences on all theoretical and practical issues arising in telecommunications, electronics, signal processing and applications, information technologies, energy systems and etc.

As always, many people have contributed to this issue including all authors, the anonymous reviewers, the Associate Editors, the Editorial Advisory Board and, of course, our outstanding production team. We wish to acknowledge each individual's contributions, and to also express our gratitude for everyone's commitment. We hope you enjoy this first issue of The Journal of CIEES.

We have had the privilege to publish truly outstanding, quality manuscripts that, judging by the response of our reviewers and readers, are important contributions to their respective fields of research. We want to take this moment to thank all of our reviewers and authors for entrusting us with their discoveries.

Editors-in-Chief:

Prof. TEODOR ILIEV, PhD
University of Ruse, Bulgaria

&

Prof. IVAYLO STOYANOV, PhD
University of Ruse, Bulgaria

MANAGING EDITORIAL BOARD

Editors-in-Chief:

Prof. Teodor ILIEV, PhD
University of Ruse
Bulgaria

Prof. Ivaylo STOYANOV, PhD
University of Ruse
Bulgaria

Associate Editor-in-Chief:

Prof. Grigor MIHAYLOV, PhD
University of Telecommunications and Post
Bulgaria

Chairman of the Editorial Advisory Board:

Prof. RENATO FILJAR, PhD
Krapina University of Applied Sciences
Croatia

Editorial Advisory Board:

Prof. Iliya ILIEV, PhD - University of Ruse, Bulgaria
Assoc. Prof. Bogdan POPA, PhD - University POLITEHNICA of Bucharest, Romania
Assoc. Prof. Gani BALBAYEV, PhD - Academy of logistics and transport, Kazakhstan
Prof. Felipe PENARANDA-FOIX, PhD - Universidad Polit cnica de Valencia, Spain
Assoc. Prof. Angel TERZIEV, PhD - Technical university of Sofia, Bulgaria
Prof. Goran DJORDJEVIC, PhD - University of Nis, Serbia
Assoc. Prof. Selahattin KOSUNALP, PhD - Bandirma Onyedi Eylul University, Turkey
Assoc. Prof. Stanimir SADINOV, PhD - Technical university of Gabrovo, Bulgaria
Assoc. Prof. Marko HORVAT, PhD - University of Zagreb, Croatia
Assist. Prof. Danijela EFNUSHEVA, PhD - University "Ss. Cyril and Methodius", Republic of North Macedonia
Lecturer Irena POPOVIC - Krapina University of Applied Sciences, Croatia

Web Design Editor:

Assoc. Prof. Yordan DOYCHINOV, PhD
University of Ruse
Bulgaria

English Language Editor:

Assoc. Prof. Diana ILIEVA, PhD
University of Telecommunications and Post
Bulgaria

The Journal of CIEES

Volume 6, Issue 1 / 2026

Print (ISSN: 2738-7283) & Online (ISSN: 2738-7291)

<http://journal.ciees.eu>

E-mail: publication@ciees.eu

Published by: CIEES Association Press, Bulgaria

Copyright: Association on Communications, Information, Electronic and Energy Systems

CONTENTS

Machine Learning Algorithm for the Influence of Space Weather Events of Geomagnetic Activity on GNSS Services	7
Lucija Žužić, Deni Klen, Darko Špoljar, and Renato Filjar	
Explainable ML-Based Time Source Arbitration for 5G Networks: Feature Engineering, Contextual Memory, and the Stability Paradox.....	16
Ivelin R. Penkov	
Assessment of Digital Infrastructure Maturity in Bulgarian Scholarly Libraries: Empirical Study of Automation and AI Adoption.....	24
Viktoriya Ivanova	
Analysis and Prediction of Life Expectancy at the Global Level and in Selected Countries.....	31
Dimitrios Kazolis, Ioanna Kranioti and Theodosia Kazoli	
Security Considerations and Techniques for Video Streaming Distribution in Home ISPs.....	38
Iliyan Iliev, Ivan Blagoev and Krassimira Stoyanova	

Machine Learning Algorithm for the Influence of Space Weather Events of Geomagnetic Activity on GNSS Services

Lucija Žužić^{1,2}, Deni Klen^{1,2}, Darko Špoljar³ and Renato Filjar^{1,2,4*}

¹ Department of Computer Engineering, Faculty of Engineering, University of Rijeka,
Vukovarska 58, 51000 Rijeka, Croatia

² Center for Artificial Intelligence and Cybersecurity, University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia

³ Department for Transport Logistics, Hrvatsko Zagorje Krapina University of Applied Sciences,
Setaliste hrvatskog narodnog preporoda 6, 49000 Krapina, Croatia

⁴ Department for Informatics, Hrvatsko Zagorje Krapina University of Applied Sciences,
Setaliste hrvatskog narodnog preporoda 6, 49000 Krapina, Croatia

*E-mail of corresponding author: renato.filjar@uniri.hr

Abstract: This study proposes a machine learning-based model that utilises the Disturbance storm-time (*Dst*) index to categorise space weather events in sub-equatorial regions and assess their impact on Global Navigation Satellite System (GNSS) Positioning, Navigation, and Timing (PNT) performance. Performance was benchmarked using various indicators derived from a confusion matrix for each *Dst* class defined in the study. Several algorithms from various families, including probabilistic methods, tree-based methods, ensemble learning methods, and a deep neural network (NN), using Long Short-Term Memory (LSTM) and bidirectional or convolutional layers, were evaluated on historical *Dst*-index-labelled GNSS data. The LSTM and convolutional-layer methods achieved the highest accuracy, but other methods had much lower inference times. The study presents a novel framework for classifying the impact of geomagnetic activity on GNSS services, thereby contributing to the development of an early-warning system for PNT degradation and to the establishment of a tailored GNSS ionospheric correction model for space weather conditions.

Keywords: Global Navigation Satellite System Positioning, Navigation, and Timing, space weather conditions, machine learning, classification model, Total Electron Content, Disturbance storm-time.

Received 10 February 2026

Accepted 18 March 2026

Published 6 April 2026

1. Introduction

The Global Navigation Satellite System (GNSS) and its Positioning, Navigation, and Timing (PNT) service have matured into an essential part of national infrastructure, public goods, and are the enablers of a vast number of emerging technologies and socio-economic applications [1, 2]. Maintaining GNSS PNT quality is crucial for the sustainable development of the modern economy and society [2, 3]. Overcoming the shortcomings and vulnerabilities of GNSS PNT poses a scientific challenge [4, 5, 6] and is necessary for a diverse range of scientists, developers, operators, regulators, and users of GNSS-based systems and services [2]. Modifications and novel approaches to PNT algorithms, information fusion, and supporting computing and communications architectures have been under development to make GNSS

PNT resilient against the natural and artificial adversarial effects [2, 6]. Those include introducing GNSS positioning environment (ambient) awareness and alignment to GNSS applications to meet their needs and demands for the PNT performance [6].

The integrity of GNSS services is increasingly threatened by adverse space weather conditions, particularly geomagnetic storms [4, 7, 8]. The morphology of these intense geomagnetic storms [6, 9] can be studied using geophysical data analysis based on geomagnetic observatory data, solar wind flux speed, and interplanetary magnetic field parameters [7, 8]. The Disturbance storm-time (*Dst*) index is a widely accepted geomagnetic activity indicator, used to evaluate the intensity of such storms [4, 7, 8]. The geomagnetic activity, including pulsations, is strongly modulated by storm conditions, local time, and magnetic latitude. Understanding the magnetosphere-ionosphere coupling, introduced and described in [10], is important for space weather prediction. Space weather, geomagnetic, and ionospheric conditions may modify the GNSS signal characteristics, thus leading to significant PNT degradation, especially in safety-critical applications



© The Author(s). Published by CIEES Association Press. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 Generic License.

such as aviation and maritime navigation. Recent developments in artificial intelligence have enabled the classification and prediction of space weather conditions using machine learning (ML) methods [10, 11, 12].

The Earth's ionosphere, a part of the Earth's atmosphere stretching from 50 km to 2000 km above the Earth's surface and comprised of rare but mostly charged atoms and molecules, is the major natural cause of GNSS PNT degradation [7]. This phenomenon especially affects most currently used GNSS receivers, which operate as single-frequency receivers and are exposed to GNSS ionospheric effects [1]. Driven by powerful and random flows of energy and particles from the Sun (space weather conditions), the ionospheric conditions define the properties of GNSS signal propagation through the Earth's atmosphere and the resulting ionospheric delay [7]. The GNSS ionospheric delay introduces errors in the GNSS PNT algorithm, which is designed to produce position, velocity, and time estimates for a GNSS receiver [1, 3]. The ionosphere affects GNSS satellite signals for position estimation by introducing signal propagation delay [1, 3]. The GNSS ionospheric delay is a stochastic variable whose value is determined by complex physical processes of space weather [1, 3]. The GNSS signal encounters a certain number of charged particles from the satellite to a mobile unit (GNSS receiver) [3]. This encounter is quantified using the Total Electron Content (*TEC*) defined by (1):

$$TEC = \int_{\text{upper ionospheric boundary}}^{\text{lower ionospheric boundary}} N(h)dh \quad (1)$$

in electrons/m² [1, 3, 7], where h denotes the height above the Earth's mean sea level in m, $N(h)$ represents the vertical ionospheric profile, the volume density of charged particles at height h , in electrons/m². For convenience, *TEC* may be expressed in *TEC* Units (TECU), with 1 TECU = 1016 electrons/m². The *TEC* dataset used in this study was derived from Receiver Independent Exchange Format (RINEX) *TEC* observations using Global Positioning System (GPS) *TEC* software by Seemala [13].

It should be noted that in the sense of (1), *TEC* is defined as a result, a consequence, of the ionospheric conditions, and not their descriptor [5, 6, 11]. The GNSS ionospheric delay [10] may be determined by derivation from the Appleton-Hartree equation, as given in (2):

$$\Delta t_{iono} = \frac{40.3}{cf^2} \int_{\text{upper ionospheric boundary}}^{\text{lower ionospheric boundary}} N(h)dh \quad (2)$$

where Δt_{iono} denotes the GNSS ionospheric delay in s, c denotes the velocity of an electromagnetic wave in vacuum in m/s, and f denotes the carrier wave frequency of the satellite signal in Hz [1, 3].

Combining (1) and (2), one can conclude the linear relation between *TEC* and Δt_{iono} , as given in [1] and (3):

$$\Delta t_{iono} = \frac{40.3}{cf^2} TEC \quad (3)$$

The GNSS ionospheric delay has been identified as a source of GNSS PNT degradation since the dawn of GNSS. GNSS systems offer various standard GNSS ionospheric delay estimation (correction) models to mitigate the deteriorating effects on GNSS PNT, such as the Klobuchar model [1, 8]. The standard ionospheric correction models are global and insufficiently flexible to update, thereby mitigating GNSS ionospheric delay and satisfying the rising GNSS PNT performance demands [1].

The development of regional and local models, such as the one presented in this paper, aims to address the issue of GNSS PNT sustainable performance under various ionospheric conditions. Understanding the long-term non-linear association between geomagnetic disturbances and GNSS positioning errors is a key step in developing advanced correction procedures to mitigate these environmental effects. The Ambient-Aware Application-Aligned (abbreviated as (AA)²) PNT has been introduced recently to account for the actual ionospheric and geomagnetic conditions in the immediate vicinity of a mobile unit (a GNSS receiver) [6]. Direct measurements of the immediate geomagnetic and ionospheric condition variables can be supplied to a machine-learning-based, adapted GNSS ionospheric correction model, thereby providing the framework for solving the single-frequency GNSS problem [6]. Previous research has identified predictors and target variables (descriptors of geomagnetic, ionospheric, and GNSS PNT conditions). Current space weather severity scales, such as those provided by the National Oceanic and Atmospheric Administration (NOAA), are based on global space weather and geomagnetic indices averaged over a given period, such as 3 hours for the global K_p index [11].

This study describes a methodology for machine learning-based classification of ionospheric conditions, which is entirely based on observations of geomagnetic indices. The model is sufficiently simple to be applied on computationally capable platforms with suitable geomagnetic field sensors, such as smartphones and connected/autonomous vehicles. The research presented acquires ambient data and examines its statistical properties. The dataset is split into training and test sets. Several candidate GNSS ionospheric delay models are developed for *Dst* data from the INTERMAGNET [14] dataset reformatted to match the *TEC* data format. The ML methods represent examples from different families, including probabilistic methods (a linear model (LM)), tree-based methods implemented in the "rpart" package, ensemble learning methods (generalised additive model boosting (GAMB), bagged classification and regression trees (BCART), and gradient boosting machine (GBM)), and deep neural network (NN) models using Long Short-Term Memory (LSTM) and bidirectional or convolutional layers. The proposed deep learning LSTM architecture is an original contribution by the authors, redesigned for multiple classes in a different research area, and represents a significant extension of the original architecture's previous application in binary peptide classification.

ML approaches in space weather classification generally fall into two categories: traditional statistical models and deep learning architectures. Traditional methods, such as LMs and tree-based models, offer high interpretability and computational efficiency but may struggle to capture complex non-linearities in geophysical data. Conversely, deep learning models, particularly Recurrent Neural Network (RNN) models and LSTM networks, excel at modelling temporal dependencies in sequential data, making them well-suited for characterising the evolution of geomagnetic storms.

A tailored set of validation methods is used to assess their performance. The optimal GNSS ionospheric delay

correction model is identified based on GNSS PNT-related objective criteria, and its performance is demonstrated in an independent case study. It is important to clarify that this framework is designed for nowcasting and classification rather than forecasting future indices. By classifying the current storm phase based on real-time indices, the system serves as a trigger for adaptive GNSS algorithms. Upon detecting an extreme or recovery phase, the PNT engine can immediately switch to a specialised ionospheric correction model or re-weight satellite observations, thereby mitigating PNT degradation during the event.

2. Method

A linear model (LM), tree-based methods from the "rpart" package, ensemble learning methods (generalised additive model boosting (GAMB), bagged classification and regression trees (BCART), and gradient boosting machine (GBM)), and custom deep neural network (NN) models using Long Short-Term Memory (LSTM) and bidirectional or convolutional layers [15, 16, 17] were tested based on their ability to classify a set of observations [18] of the geomagnetic field in *TEC*, and other predictors, into one of the scenarios of space weather conditions based on *Dst*. The methods were selected because they represent larger families [15, 16, 17] of machine learning methods [18]. Where appropriate, simpler linear models may be preferable unless deep learning models substantially improve predictive performance. This is why simpler models were benchmarked alongside computationally intensive LSTM models. The full data and code are on Zenodo (<https://doi.org/10.5281/zenodo.14502657>, accessed on 3 September 2025).

All tested models, except the LSTM deep learning models, were applied as described in the research by Kuhn for the "caret" package [19], which was used in R [20]. The deep learning models were implemented in Python version 3.10.13 using the "keras" library version 2.10.0 to replicate the authors' previous research, which used a similar model architecture in the domain of peptide self-assembly [21].

Cross-validation can be used to detect and potentially reduce overfitting problems with statistical methods [21]. The "trainControl" function was used to tune the hyperparameters of the non-deep learning models implemented in the R environment [20] using the "caret" package [19] in k-fold cross-validation with 10 folds and repeated 10 times.

To measure inference time and training time, the experiment was run on the 6th Generation Intel(R) Core(TM) i5-6300HQ Central Processing Unit (CPU) of a personal computer, running at a base speed of 2.3 GHz, instead of a Graphics Processing Unit (GPU), to better estimate the execution time for mobile devices and built-in systems using GNSS PNT that have significantly lower computational capabilities.

An algorithm (block diagram) of the method is presented in Figure 1.

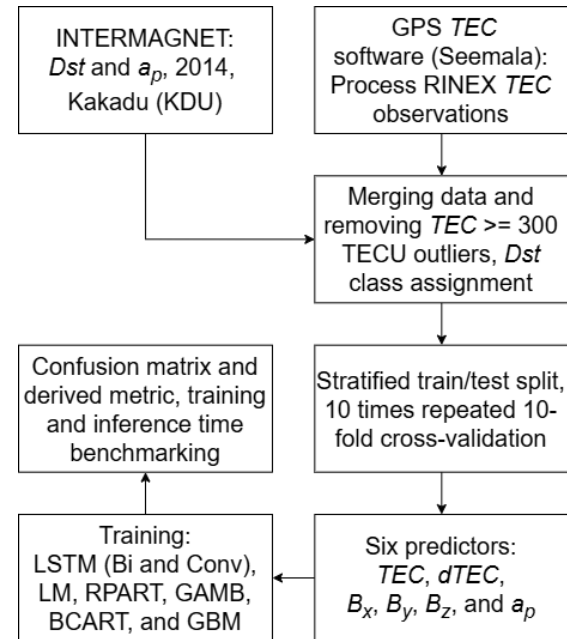


Fig. 1. The full method workflow from data processing to ML model training and testing.

GPS *TEC* software by Seemala was used to obtain RINEX *TEC* observations [13], and this data was merged with the INTERMAGNET [14] dataset, which included *Dst* and a_p data from 2014 for a measuring station maintained by Geoscience Australia in Kakadu (KDU). The $TEC \geq 300$ TECU outliers were removed before *Dst* classes were determined. The training and testing datasets were stratified by class, and the models were evaluated using 10-fold cross-validation (repeated 10 times) [18, 19, 20]. Six predictors were used in the LSTM (bidirectional (Bi) and convolutional (Conv)), LM, RPART, GAMB, BCART, and GBM models for training [15, 16, 17]. Results evaluation included confusion matrix performance indicators and comparisons of inference and training times.

2.1. Linear models

The linear model (LM) assumes linearity [15, 19, 20]. A substantial reduction in the complexity of the related statistical theory is possible by using least squares analysis, such as the minimum of a sum of squares function [16, 17, 18].

2.2. Tree models

A decision tree (DT) is used for supervised learning in statistics and machine learning [15, 16, 17]. A tree is recursively partitioned by dividing the original set (the root node) into subsets that form descendants (successors), using classification rules based on features [18]. Random forests (RFs) are ensemble models that implement bagging over a set of DT classifiers. This effectively decreases overfitting. Multiple trees with different feature subsets [22] and training samples are created to avoid depending on a single tree. The final decision is made on the majority classification of these trees.

2.3. Generalised additive model boosting

In a generalised additive model (GAM), the response variable depends linearly on the smooth functions of one or more predictor variables, combining the properties of generalised linear models and additive models. GAMs can be interpreted as the discriminative generalisation of the naive Bayes generative model [15, 16, 17]. Parametric, nonparametric, or semiparametric function forms can be used to improve fit [18]. These methods include the risk of losing interpretability [15, 16, 17]. GAM Boosting (GAMB) fits a generalised additive model using likelihood-based boosting [18].

2.4. Bagged tree models

Bootstrap aggregating, or bagging, is an ensemble meta-algorithm that increases the stability and accuracy of ML classification and regression algorithms by combining the predictions of randomly generated training sets [15, 16, 17]. Many weak learners typically aggregate to outperform a single learner across the entire set and tend to overfit less [18]. Training for individual weak learners can be performed in parallel, as each separate bootstrap can be processed independently.

2.5. Gradient boosting machines

A gradient boosting machine (GBM) is an algorithm based on boosting in a functional space, where the target is pseudo-residuals rather than residuals, as in traditional boosting [15, 16, 17]. The idea of gradient boosting originated from the observation that boosting can be interpreted as an optimisation algorithm [18].

2.6. LSTM bidirectional and convolutional models

The LSTM architecture conceived by the authors for binary peptide classification [21] has been successfully adapted and extended for *Dst*-index multiclass classification. Inspired by peptide [21] feature scaling, the variable values were again scaled to the [0, 1] range, facilitating gradient flow and yielding a model with superior performance [23]. Since the number of predictors was uniform, a masking layer and padding for processing varying peptide sequence lengths were unnecessary. Unlike in previous work on peptide classification [21], there were no sequences of multiple values for the same feature, such as AP (Aggregation Propensity) of amino acids, dipeptides, and tripeptides, so a concatenation layer was omitted. One version of the tested models applied two LSTM layers with 5 units to the input data, with the first layer being bidirectional. These submodels also encompassed a dense layer with 64 units, as obtained by hyperparameter optimisation in previous work [21], followed by the scaled exponential linear unit (SELU) activation. Other tested models utilised two one-dimensional convolutional layers, with 5 filters and a kernel size of 4, as determined through hyperparameter tuning in the preceding experiments on peptides [21]. The number of filters was inspired by previous research on therapeutic peptides [24]. The training was limited to 5 epochs [25], and the batch size was set to 600 for the fastest operation and smoother gradients. The initial learning rate was set to 0.01. Only the model with the lowest training

loss was saved. A 50% dropout regularisation was applied to the final layer of all submodels to prevent overfitting. The sigmoid activation function was applied to the final layer of each model to yield a value between 0 and 1, representing the probability that an input sequence exhibits self-assembly in the original model. In the *Dst* model, this value is converted to the original variable range, with 0 representing the minimum and 1 representing the maximum. Values are then rounded to the nearest integer to obtain the class number.

2.7. Confusion matrix and statistical test specifications

All tested candidate models were compared to the ground truth. A confusion matrix and derived metrics were used, including the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications, accuracy, true positive rate (TPR) or sensitivity, true negative rate (TNR) or specificity, *F1* score, balanced accuracy (BA), positive predictive value (PPV), and negative predictive value (NPV) [15, 16, 17]. These metrics [18] and the 95% confidence interval (CI) were calculated using the "epi.tests" function from the "epiR" package developed for R [20]. The parameters "conf.level" and "digits" were 0.95 and 4, respectively. The inputs were the TP, TN, FP, and FN values shown in confusion matrix heatmaps in Figure 2 in further text for each tested model.

3. Data

Dynamic space weather conditions, such as solar activity and geomagnetic storms, can impact GNSS PNT performance and affect high-frequency GPS signals passing through the ionosphere, motivating research on error modelling. Geomagnetic storms degrade signal quality by altering the variability of the Global Electric Current (GEC). The ionosphere may show changes related to location, geomagnetic and solar activity, sunspots, local time, seasonality, thunderstorms, nuclear experiments, earthquakes, and other phenomena. This study focuses on parameters describing disturbances of the Earth's geomagnetic field, most importantly a_p -indices derived from K_p -indices, which are calculated using K -indices, the Total Electron Content (*TEC*), the standard deviation of the Total Electron Content (*dTEC*), and Disturbance storm-time (*Dst*). Incorporating parameters such as the K_p -indices and a_p -indices, which provide global measures of geomagnetic activity, alongside local *TEC* and *Dst*-index values, enables a more detailed assessment of the space environment and its potential effects on GNSS signals. Values of a_p , *TEC*, *dTEC*, and *Dst* were used with the B_x , B_y , and B_z components of the Earth's magnetic field to train machine-learning models. The *dTEC* was analysed based on all measurements in each *Dst* class and derived from the RINEX observations within the hourly window.

Multiple *Dst*-dependent classes were predefined using theoretical knowledge. Statistical analysis of the data confirmed that the distributions of other variables differ across different *Dst* ranges, supporting the validity of the classification. The study assumes that the dependent output variable, the *Dst* class, can be predicted from the independent input variables. *TEC* data were obtained using GPS *TEC* software by Seemala to process RINEX *TEC*

observations [13]. The INTERMAGNET [14] dataset contains Dst and a_p data from 2014 for a measuring station maintained by Geoscience Australia in Kakadu, referred to as KDU in the database, at 12.69 degrees south latitude and 132.47 degrees east longitude, near Darwin, Northern Territory, Australia. The two datasets are merged based on location, year, month, day, and time of day (hours and minutes).

3.1. Magnetic field indices

The Earth's magnetic field resembles that of a bar magnet. However, plasma emanating from the solar corona and the Sun's domain both influence the interplanetary magnetic field. The B_x , B_y , and B_z vectors represent the components of the interplanetary magnetic field. B_x and B_y are parallel to the plane of orbits, and the third component B_z is perpendicular.

3.2. Geomagnetic storm indices

The geomagnetic storm K -index is an integer from 0 to 9 measuring disturbances in global geomagnetic activity. The maximum positive and negative fluctuations of the horizontal components of the Earth's magnetic field, B_x and B_y , over 3 hours relative to a quiet day are added to determine the total maximum fluctuation. Each observatory uses different threshold values to convert the maximum nT (nano-Tesla) fluctuation to a K -index value. The thresholds for each observatory are adjusted so that the historical occurrence rates for each K -index value are similar across all observatories. Observatories with a lower geomagnetic latitude use a lower fluctuation in B_x and B_y to achieve each K -index value.

3.3. Planetary geomagnetic storm indices

The planetary geomagnetic storm K_p -index is derived from 3-hour-based K -indices from 13 magnetometer stations between 44 and 60 degrees of north and south latitude. Hourly geomagnetic storm index data are available on the National Aeronautics and Space Administration (NASA) Goddard Space Physics Data Facility (SPDF) web pages. The scale values of the K_p -index are determined by the change of the geomagnetic field and the geomagnetic storm effect in nT. The official planetary K_p -index is a weighted average of multiple observatories. When K -index data is not available in real-time, operators such as the National Oceanic and Atmospheric Administration (NOAA) Space Weather Prediction Center (SWPC) calculate near-real-time estimates of the K_p -index [26].

3.4. Equivalent three-hourly range geomagnetic storm indices

The A -index represents a daily average level of magnetic activity. Because the relationship between the K -index and magnetometer fluctuations is nonlinear, the K -index values are not directly used for calculating average values. Each K -index or K_p -index is converted to the "equivalent three-hourly range" a -index or a_p -index, which uses a linear scale. An average of 8 a -indices (lowercase) is used as the daily A -index (uppercase).

3.5. Disturbance storm-time

Disturbance storm-time (Dst), also known as the geomagnetic activity Dst Index, represents an averaged measure of geomagnetic storm intensity in the Earth's subequatorial region. The Dst -index is obtained by post-processing, with the final version often published months after experimental observations were collected. It clearly shows the progression of geomagnetic disturbances at various levels and, consequently, provides invaluable input for assessing GNSS PNT performance degradation due to space weather events. The Dst -index is a geomagnetic indicator of magnetic flux changes derived from measurements taken by a network of ground-based magnetometer stations near the magnetic equator, which continuously monitor B_x and B_y , the horizontal components of Earth's magnetic field [27]. The Dst -index describes ring currents that form above the sub-equatorial region and affect the ionospheric regions in mid-latitudes. To calculate the Dst -index, variations in B_x and B_y , the horizontal magnetic field, are obtained from multiple stations and averaged. The average is subtracted from a baseline value representing the quiet-time magnetic field. The resulting Dst value in nT measures the intensity of geomagnetic disturbances, with increasingly negative values indicating stronger geomagnetic storms. The Dst -index measurements (as an hourly average) were evaluated and published on a web interface through a collaboration between NASA Goddard SPDF and the Geomagnetism and Space Magnetism Data Analysis Center of the Institute of Science, Kyoto University, Japan. Loewe and Pröls [28] classified magnetic activity Dst -indices into 5 storm classes in 1997. Gonzalez et al. [29] used 3 groups for the same data in 1994, similar to Kamide et al. in 1998 [30], Rozhnoi et al. in 2004 [31], and Contadakis et al. in 2012 [32].

3.6. Training, testing, and validation dataset

The original set of observations has been split into training and test subsets while sustaining proportions of data classes for geomagnetic disturbances (quiet/normal geomagnetic conditions, the positive phase of a geomagnetic storm, deep negative depression of a geomagnetic storm, and the negative recovery phase of a geomagnetic storm). Classes are based on Dst and differential Dst time series. Table 1 lists the Dst class ranges used in this study, the total number of samples per class, and the numbers used for testing and training.

Table 1. Dst -based classification rules used in this study, the total number of samples in each class, and the number of samples used for testing and training. The upper limit is excluded, while the lower limit is included.

Dst	Storm phase	Total	Test	Train
≥ 15	positive (P)	8,360	1,672	6,688
$[-20, 15>$	normal (N)	379,167	75,832	303,335
$[-55, -20>$	recovery (R)	115,856	23,168	92,688
$[-85, -55>$	through (T)	5,666	1,132	4,534
< -85	extreme (E)	832	164	668
Any	Any	509,881	101,968	407,913

To address the differing temporal resolutions of the input variables, such as hourly *Dst*, daily *A*-index, and 30-second RINEX data, all datasets were synchronised to the cadence of the high-frequency RINEX TEC measurements. The *Dst* index values and the daily *A*-index can be repeated as needed.

522,297 represents the total count of TEC observations available in the dataset. Of 522,297 original samples, 12,416 were identified as outliers in exploratory statistical analysis, as their *TEC* is 300 TECU or more. *TEC* values exceeding 300 TECU are usually due to measurement errors or errors in the *TEC* estimation process. 509,881 samples are used for training and testing, with an 80:20 split. The division was stratified using the "createDataPartition" function from the R library [20] "caret" [19]. The training dataset was further divided into training and validation sets during 10-fold cross-validation, with 10 repeats, for all models, using the "trainControl" function from the "caret" package [19]. Samples were split into 5 class ranges, based on *Dst* values derived from theoretical knowledge of different storm phases. A similar class division was proposed by Loewe and Pröls [28]. It is evident from Table 1 that the normal (N) and recovery (R) classes are more common than those with very high or very low *Dst* values, which impacts model performance. The values "scale" and "centre" were used in this study in the "preProcess" parameter for the train function from the R package [20] "caret" [19]. The option "center" subtracts the mean of each feature, while "scale" divides by the standard deviation.

4. Results

Candidate models were evaluated to determine the optimal method and set of predictors to be included in the final model for application in real-world settings. Research results are presented with the following set of six predictors: the Total Electron Content (*TEC*), the standard deviation of Total Electron Content (*dTEC*), B_x , B_y , B_z , and a_p . Theoretical studies show that geomagnetic indices B_x , B_y , and B_z affect *Dst*, and, in turn, *TEC*. This is why geomagnetic indices (B_x , B_y , and B_z), *TEC*, *dTEC*, and a_p were used as predictors in the final model.

The evaluated methods using the full set of predictors are abbreviated as follows in further text: the LSTM deep learning method with bidirectional layers (Bi), the LSTM deep learning method with convolutional layers (Conv), the linear model (LM), the classification and regression trees implemented in the "rpart" package (RPART), the generalised additive model boosting (GAMB) method, the bagged classification and regression trees (BCART), and the gradient boosting machine (GBM) method.

For multiple classes, the values for each metric were calculated by a "one versus all" approach, viewing each

class as positive and all others as the opposite, and then producing the arithmetic average across all classes, presented in Figure 3 as a bar chart and in Figure 2 as a confusion matrix heatmap.

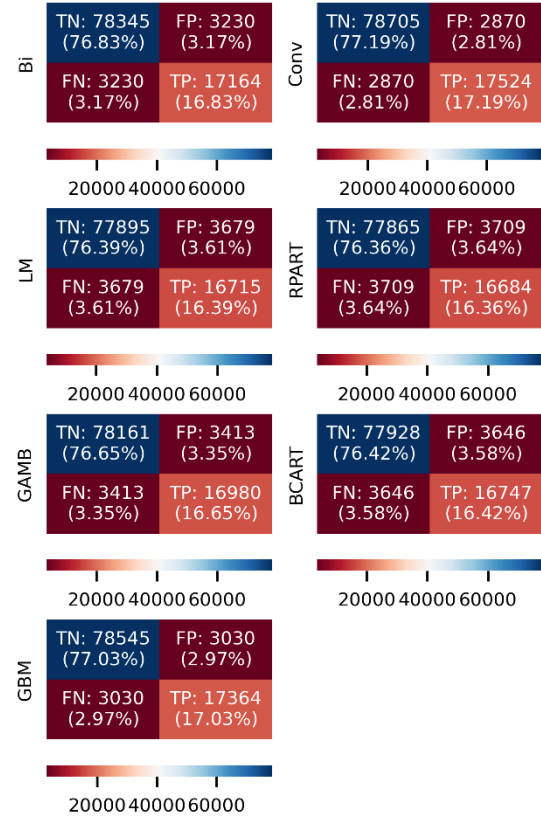


Fig. 2. The confusion matrix heatmap with the number and percentage (%) of TN, FP, FN, and TP classifications for the arithmetic average of all the *Dst*-index classes.

The Conv model had the highest numbers and percentages (%) of TP and TN classifications and the lowest numbers and percentages (%) of FP and FN classifications, while the opposite is true for the RPART model. The Conv model had the highest TPR, TNR, PPV, NPV, accuracy, BA, and *F1* score, while the RPART model had the worst performance across the *Dst*-index classes and is thus unsuitable for this problem. The arithmetic average across all *Dst*-index classes clearly positions the Conv model as the most accurate, with the highest scores across all metrics. In contrast, the RPART model achieved the lowest scores across all metrics, confirming that it cannot be used for the application intended in this study.

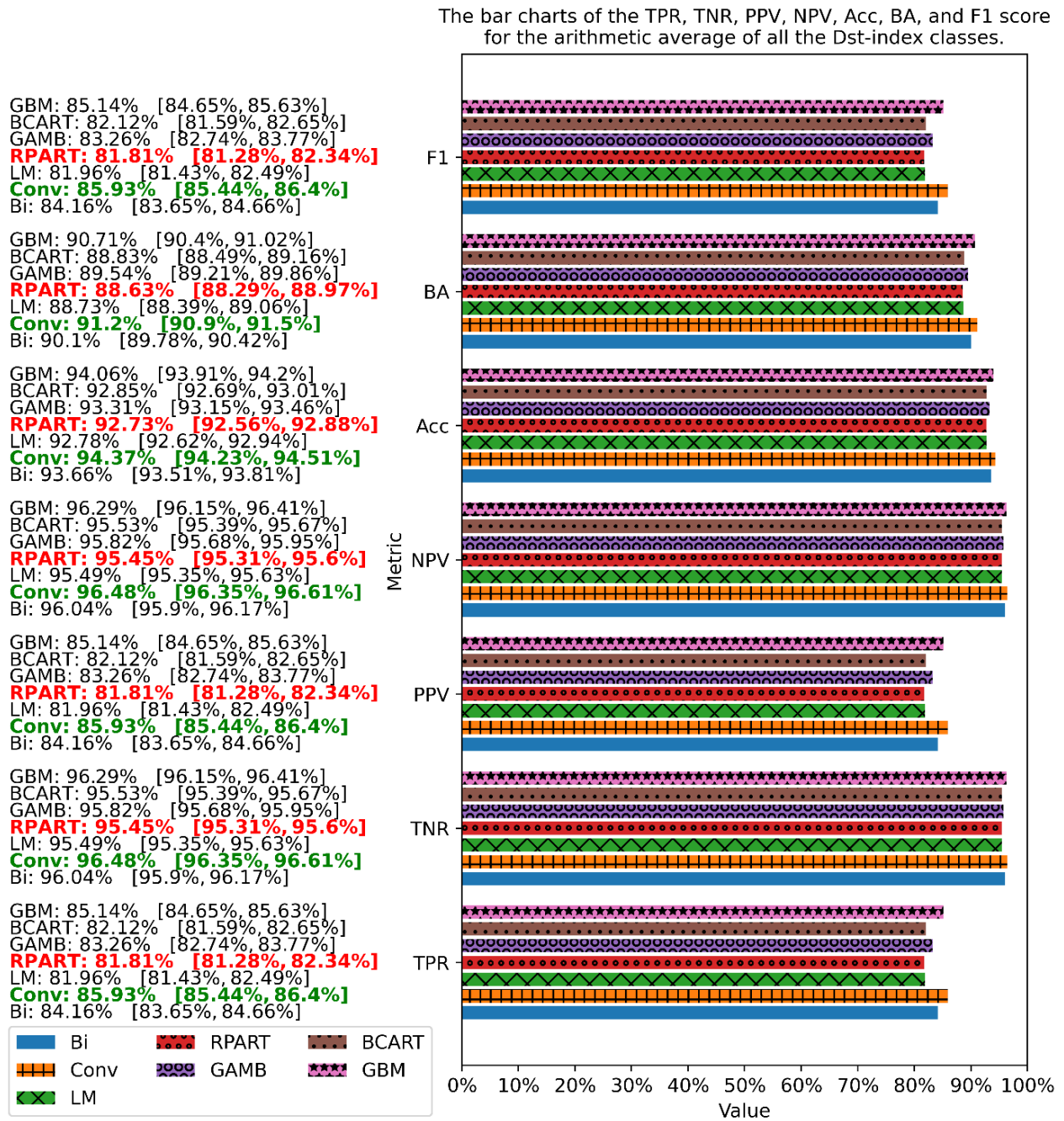


Fig. 3. The bar charts of the TPR, TNR, PPV, NPV, Acc, BA, and F1 score for the arithmetic average of all the Dst-index classes. The 95% CI is in square brackets. The highest scores for each metric are marked in bold and green, and the lowest scores are bold and red.

Table 2. The total training and testing time for all methods, with the lowest values marked in bold.

Time	Bi	Conv	LM	RPART	GAMB	BCART	GBM
Train	35.68 s	61.31 s	3.8 s	12.31 s	36.44 s	1.22 s	1.39 s
Test	2.45 s	2.23 s	0.03 s	0.77 s	7.14 s	1.66 s	1.32 s

The interpretation of the results must also weigh statistical rigour against computational feasibility. The total training and testing time for each candidate model is presented in Table 2. The total model training and testing time, in seconds (s), for all samples, utilising the R "system.time" function [20], across all candidate models, is displayed in Table 2.

The LM method required the least testing time because of its simplicity, and the BCART method required the least training time. The Bi and Conv methods have the highest computational demands due to deep learning implementation. The Conv model required 61.31 s for training (on the test setup) compared to just 1.22 s for BCART. While the Conv model is computationally

intensive, the faster models (LM and RPART) are disqualified because they cannot detect the safety-critical extreme phases of ionospheric storms.

5. Conclusion

This work introduced and validated a novel machine learning framework for classifying geomagnetic storm phases using the *Dst*-index and related ionospheric and geomagnetic variables, with the overarching goal of supporting GNSS PNT integrity in the face of space weather disturbances. This work contributes toward enabling (AA)² PNT systems through models deployable on smartphone-class platforms. This study compares traditional statistical models, tree-based ensembles, and deep learning networks. Five discrete *Dst*-based storm phases are identified and used as labels consistent with prior work. The findings show that deep learning architectures, such as convolutional and bidirectional LSTM models, outperform the competition, particularly for identifying rare and extreme space weather events that are critical to PNT reliability. These models successfully captured complex, non-linear relationships between GNSS-relevant features (*TEC*, *dTEC*, *B_x*, *B_y*, *B_z*, and *a_p*) and *Dst*-index storm classifications, outperforming simpler methods that failed to generalise to minority classes. The accuracy benefits come with the caveat of a higher computational cost. The presented research offers guidelines for selecting appropriate models based on deployment context, including cloud-based systems and embedded platforms. This work builds on previous contributions that warned of the need for a local, context-dependent GNSS model. The developed methodology for classifying space weather events represents a foundation for novel ambient-aware correction approaches to improve GNSS performance in the next generation of systems. The use extends to aviation, autonomous vehicles, maritime navigation, and emergency response. Future research directions include expanding the dataset to include more locations and dates, implementing more complex resampling methods, applying anomaly detection, or adopting hybrid ensemble strategies.

References

- [1] Parkinson, B.W. & Spilker Jr., J.J. (1996). *Global Positioning System: Theory and Applications, Volume I*. American Institute of Aeronautics and Astronautics, Reston, VA, USA.
- [2] Filjar, R., Damas, M.C., & Iliev, T.B. (2020). Resilient Satellite Navigation Empowers Modern Science, Economy, and Society, *CIEES 2020*, Borovets, Bulgaria, 2020, *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1032, article no. 012001, doi: 10.1088/1757-899X/1032/1/012001.
- [3] Schaer, S. (1999). *Mapping and predicting the Earth's ionosphere using the Global Positioning System*. Institut für Geodäsie und Photogrammetrie, Eidg. Technische Hochschule, Zürich, Switzerland.
- [4] Oxley, A. (2017). *Uncertainties in GPS Positioning: A Mathematical Discourse*. Academic Press/Elsevier, London, UK.
- [5] El Yousoufi, H., Emran, A. & Filjar, R. (2024). Identification of GPS PNT degradation using analysis of non-linear properties of GPS-derived TEC time series, *ION ITM 2024*, Long Beach, CA, pp. 135-145, doi: 10.33012/2024.19527.
- [6] Filjar, R. (2022). An application-centred resilient GNSS position estimation algorithm based on positioning environment conditions awareness, *ION ITM 2022*, Long Beach, CA, pp. 1123-1136, doi: 10.33012/2022.18247.
- [7] Davies, K. (1990). *Ionospheric Radio*. Peter Peregrinus Ltd., London, UK.
- [8] Klobuchar, J.A. (1987). Ionospheric time-delay algorithm for single-frequency GPS users, *IEEE Trans. on Aerosp. and Electronic Systems*, vol. 23, pp. 325-331, doi: 10.1109/TAES.1987.310829.
- [9] Žužić, L., Škrlić, L., Nešković, A. & Filjar R. (2025). Characterisation of GPS Horizontal Positioning Errors and *Dst* Using Recurrence Plot Analysis in Sub-Equatorial Ionospheric Conditions, *Urban Sci.*, vol. 9, no. 11, article no. 451, doi: 10.3390/urbansci9110451.
- [10] Filić, M. & Filjar, R. (2018). Modelling the Relation between GNSS Positioning Performance Degradation, and Space Weather and Ionospheric Conditions using RReliefF Features Selection, *31st International Technical Meeting ION GNSS+ 2018*, Miami, FL, pp. 1999-2006, doi: 10.33012/2018.16016.
- [11] Filjar, R. (2022). A contribution to short-term rapidly developing geomagnetic storm classification for GNSS ionospheric effects mitigation model development, *ICEASE 2021 Conference*, Islamabad, Pakistan, doi: 10.1109/ICASE54940.2021.9904168.
- [12] Filjar, R. (2022). On performance assessment of machine learning-based GNSS ionospheric delay correction model, *International Workshop on Machine Learning for Space Weather: Fundamentals, Tools, and Future Prospects*. International Centre for Theoretical Physics (ICTP), Buenos Aires, Argentina.
- [13] Seemala, G. (2024). GPS-TEC analysis program (version 3.5). <https://seemala.blogspot.com/> (accessed on 3 September 2025).
- [14] INTERMAGNET. (2024). International Real-time Magnetic Observatory Network. <https://intermagnet.org/> (accessed on 3 September 2025).
- [15] Kuhn, M., Johnson, K. (2013). *Applied Predictive Modeling*. Springer Nature, Cham, Switzerland.
- [16] Biecek, P. & Burzykowski, T. (2021). *Explanatory Model Analysis*. CRC Press, Boca Raton, FL.
- [17] Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Cambridge, UK.
- [18] Boehmke, B. & Greenwell, B. (2020). *Hands-On Machine Learning with R*. CRC Press, Boca Raton, FL.
- [19] Kuhn, M. (2007). caret: Classification and Regression Training, doi: 10.32614/cran.package.caret (accessed on 3 September 2025).
- [20] R Core Team (2025). R: A Language and Environment for Statistical Computing, version 4.5.1. <https://www.r-project.org> (accessed on 3 September 2025).
- [21] Njirjak, M., Žužić, L., Babić, M., Janković, P., Otović, E., Kalafatović, D. & Mauša, G. (2024). Reshaping the discovery of self-assembling peptides with generative AI guided by hybrid deep learning, *Nat. Mach. Intell.*, vol. 6, pp. 1487-1500, doi: 10.1038/s42256-024-00928-1.
- [22] Fennell, P.G., Zuo, Z. & Lerman, K. (2019). Predicting and explaining behavioral data with structured feature space decomposition, *EPJ Data Science*, vol. 8, pp. 2-3, doi: 10.1140/epjds/s13688-019-0201-0.
- [23] Singh, D. & Singh, B. (2020). Investigating the impact of data normalization on classification performance, *Appl. Soft Computing*, vol. 97, article no. 105524, doi: 10.1016/j.asoc.2019.105524.
- [24] Otović, E., Njirjak, M., Kalafatović, D. & Mauša, G. (2022). Sequential Properties Representation Scheme for Recurrent Neural Network-Based Prediction of Therapeutic Peptides, *J. of Chemical Inf. and Modeling*, vol. 62, pp. 2961-2972, doi: 10.1021/acs.jcim.2c00526.
- [25] Altché, F. & de La Fortelle, A. (2017). An LSTM network for highway trajectory prediction, *2017 IEEE 20th International*

- Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, pp. 1929-1958, doi: 10.5555/2627435.2670313.
- [26] Matzka, J., Bronkalla, O., Tornow, K., Elger, K. & Stolle, C. (2021). Geomagnetic Kp index, doi: 10.5880/Kp.0001 (accessed on 3 September 2025).
- [27] Zolesi, B. & Cander, Lj. R. (2014). Ionospheric prediction and forecasting, Springer Nature, London, England.
- [28] Loewe, C. A. & Prölss, G. W. (1997). Classification and mean behavior of magnetic storms, *Journal of Geophysical Research: Space Physics*, vol. 102, no. A7, pp. 14209-14213, doi: 10.1029/96JA04020.
- [29] Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T. & Vasyliunas, V. M. (1994). What is a geomagnetic storm?, *Journal of Geophysical Research: Space Physics*, vol. 99, no. A4, pp. 5771-5792, doi: 10.1029/93JA02867.
- [30] Kamide, Y., Yokoyama, N., Gonzalez, W., Tsurutani, B. T., Daglis, I. A., Brekke, A. & Masuda, S. (1998). Two-step development of geomagnetic storms, *Journal of Geophysical Research: Space Physics*, vol. 103, no. A4, pp. 6917-6921, doi: 10.1029/97JA03337.
- [31] Rozhnoi, A., Solovieva, M. S., Molchanov, O. A. & Hayakawa, M. (2004). Middle latitude LF (40 kHz) phase variations associated with earthquakes for quiet and disturbed geomagnetic conditions, *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 29, no. 4-9, pp. 589-598, doi: 10.1016/j.pce.2003.08.061.
- [32] Contadakis, M. E., Arabelos, D. N., Pikridas, C. & Spatalas, S. D. (2012). Total electron content variations over southern Europe before and during the M 6.3 Abruzzo earthquake of April 6, 2009, *Annals of geophysics*, vol. 55, no. 1, doi: 10.4401/ag-5322.

Explainable ML-Based Time Source Arbitration for 5G Networks: Feature Engineering, Contextual Memory, and the Stability Paradox

Ivelin R. Penkov

Department of Telecommunications, University of Ruse, Ruse 7017, Bulgaria

ipenkov@uni-ruse.bg

Abstract: The Best Master Clock Algorithm (BMCA) of IEEE 1588 selects timing sources without memory, measured quality awareness, or explainability—three properties increasingly demanded by 5G operators managing critical synchronization infrastructure. This paper presents a feature engineering and classification methodology addressing all three limitations through: (1) a multi-scale rolling statistics framework transforming raw PTP telemetry into semantically rich features capturing anomalies, stability indicators, and drift dynamics; (2) a recursive memory mechanism (prev_decision) implementing learned adaptive hysteresis via Bayesian conditional probability, eliminating flapping without manual threshold tuning; and (3) full decision explainability through Random Forest feature importance analysis (MDI). Experimental results reveal a counterintuitive stability paradox: under high-PDV conditions, the algorithm performing 1,318 source switches achieves zero ITU-T G.8271.1 mask violations, while the algorithm performing only 2 switches records 209 violations. Feature importance analysis confirms that rolling standard deviation (σ_{TE}) and previous decision (x_{prev}) dominate in adversarial scenarios, validating the hypothesis that contextual memory is the missing component in conventional synchronization management. Sensitivity analysis across oscillator classes demonstrates that ML-based arbitration with inexpensive TCXO hardware outperforms standard BMCA with rubidium oscillators, proving that algorithmic intelligence can substitute for hardware investment.

Received 2 February 2026

Accepted 15 March 2026

Published 7 April 2026

Keywords: Explainable AI, Feature Engineering, Random Forest, Synchronization, BMCA, Flapping, Hysteresis, 5G, PTP, Oscillator Sensitivity.

1. Introduction

The management of time synchronization in 5G networks is a critical infrastructure function: timing errors exceeding $\pm 1.5 \mu s$ cause TDD frame misalignment, inter-slot interference, and cell outages [1]. The Best Master Clock Algorithm (BMCA) of IEEE 1588 [2] has served as the industry standard for source selection, but its deterministic, memoryless design creates three operational problems that network operators increasingly identify as untenable.

First, BMCA lacks measured quality awareness. It evaluates sources based on self-declared clockClass, accuracy, and variance fields in PTP Announce messages, without verifying these declarations against actual observed timing performance. A GNSS receiver under slow-drift spoofing attack continues to declare clockClass 6 while its timing signal progressively degrades [3]. Second, BMCA has no decision memory. Each evaluation cycle is independent, creating the well-documented flapping problem: when two sources have similar quality,

small noise fluctuations cause rapid, oscillatory switching that introduces transient phase errors of 50–200 ns at each transition point [4]. Industry practice addresses this through manually configured hysteresis timers and wait-to-restore intervals, but these static parameters cannot adapt to varying network conditions and require per-site tuning.

Third, BMCA provides no explainability. When a source switch occurs, the operator can determine what the new source is, but not why it was selected or what conditions triggered the change. In an era of increasing regulatory scrutiny of critical infrastructure decisions, and as telecommunications operators face audit requirements for network management actions, the inability to trace synchronization management decisions is a growing concern [5].

This paper presents a feature engineering and classification methodology that addresses all three limitations through three innovations. Key contributions: (1) multi-scale rolling statistics with mathematical formalization of the sensitivity-robustness trade-off; (2) recursive prev_decision implementing adaptive hysteresis



© The Author(s). Published by CIEES Association Press. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 Generic License.

$$P(y_t = A \mid x_t, y_{(t-1)} = A) > P(y_t = A \mid x_t, y_{(t-1)} = B) \quad (1)$$

Second, we provide a quantitative analysis of the "stability paradox," demonstrating counter-intuitively that highly dynamic adaptation (1,318 micro-switches) achieves zero mask violations, whereas traditional static persistence (only 2 switches) results in 209 violations under identical high-PDV conditions. Finally, Mean Decrease in Impurity (MDI) analysis confirms that statistical stability metrics (σ_{TE}) and learned inertia (x_{prev}) dominate feature importance in adversarial scenarios, validating the architectural design.

2. Related work

Explainability in telecom ML is addressed by Zhang et al. [10] who demonstrated Random Forest viability for anomaly detection with interpretable decision boundaries. Chandramouli et al. [5] identified XAI as essential for 6G timing services, where regulatory audit of network management decisions is increasingly required. The flapping problem in synchronization is documented by Girela-López et al. [6], with static wait-to-restore (WTR) timers as the standard but inflexible mitigation. Feature importance methods include Mean Decrease in Impurity (MDI, used here for computational efficiency), permutation importance (model-agnostic but slower), and SHAP values (theoretically grounded but computationally prohibitive for real-time telecom applications). The author's prior work [11] introduced trust-score-based control. No prior work combines learned contextual memory ($prev_decision$) with explainable multi-scale feature engineering for synchronization arbitration.

3. The flapping problem in synchronization

The flapping problem occurs when the BMCA decision boundary lies in a region of high feature-space density (Fig. 1). Let $f: \mathcal{R}^M \rightarrow \{0,1\}$ be the classifier's decision function. In the boundary zone between two source classes (e.g., GNSS and PTP with similar quality), a small noise perturbation ε can invert the decision:

$$f(x + \varepsilon) \neq f(x - \varepsilon) \quad (2)$$

While mathematically valid (the classifier seeks the local optimum of the loss function), this behavior is operationally catastrophic: each source switch generates a phase transient of 50–200 ns depending on the servo characteristics, loads the telecom management plane with configuration change notifications, and—when occurring at high frequency—produces cumulative timing degradation that can exceed the error budget allocated to the transport segment [4].

The conventional solution is to introduce hysteresis through configurable wait-to-restore timers: after selecting a source, the system must wait T_{wtr} seconds before considering a switch, even if a better source appears. This approach has two fundamental weaknesses. First, T_{wtr} is a static parameter that cannot adapt to the severity of the detected anomaly—the same timer applies whether the new source is marginally better or categorically superior. Second, its optimal value depends on network conditions (PDV levels, number of available sources, oscillator quality) that vary across deployment sites and over time, requiring expert tuning during commissioning and periodic re-evaluation.

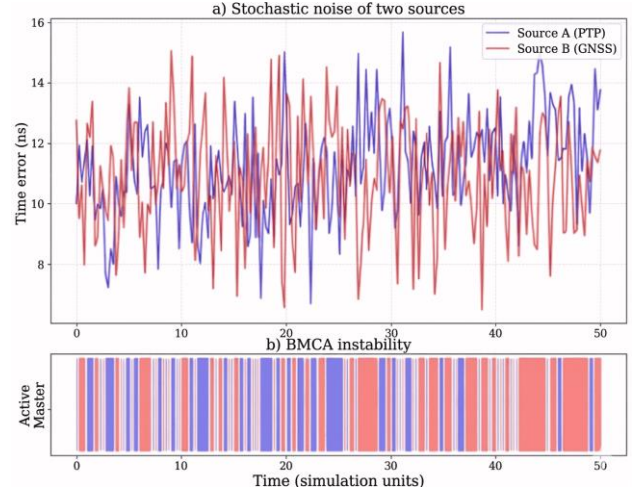


Fig. 1. Illustration of the flapping problem. When two sources have similar quality metrics, small noise variations cause rapid oscillatory switching, each introducing a phase transient that contributes to timing budget erosion.

A concrete example illustrates the dilemma: with $T_{wtr} = 60$ s, a 30 s PDV burst goes entirely undetected—the timer prevents the system from reacting to a genuine anomaly. With $T_{wtr} = 5$ s, 47 false switches occur under nominal conditions over 3600 s due to noise-induced boundary crossing. AMSA's learned hysteresis achieves zero false switches nominally while responding to genuine anomalies within one computational cycle (10 ms).

4. Multi-scale feature engineering

The proposed approach transforms raw PTP telemetry—instantaneous time error $TE(t)$, MTIE measurements, and PDV statistics—into a semantically enriched feature vector through three groups of transformations designed to provide the classifier with complete information about the current state, recent history, and dynamic trajectory of each timing source.

A. Z-Score normalization

Timing metrics span several orders of magnitude: TE ranges from 1 ns (ideal laboratory) to 500,000 ns (failed source); PDV from 10 ns (low-load fiber) to 3,000 ns (congested IP/MPLS); MTIE from 50 ns to 500,000 ns depending on observation interval. While Random Forests are theoretically invariant to monotonic feature scaling (each tree split depends only on ordering), empirical results demonstrate that Z-score normalization improves training convergence and model interpretability:

$$x'_{norm} = \frac{x - \mu_{global}}{\sigma_{global}} \quad (3)$$

Where μ_{global} and σ_{global} are computed from a reference dataset representing nominal network operation during the initialization period. This transformation maps all features to a common scale of standard deviations from the mean, effectively projecting physically heterogeneous quantities into a unified anomaly space. Critically, a value $|x'_{norm}| > 3$ constitutes a statistically significant anomaly (3σ rule) regardless of the underlying physical source. This property is essential for out-of-distribution robustness: for example, a catastrophic GNSS failure producing a TE of 442,513 ns maps to approximately

$+500\sigma$. This places the data point unambiguously in the rejection zone, ensuring correct classification even if the training set never contained such extreme values.

B. Rolling statistics (dynamic features)

Since Random Forest classifiers operate on independent feature vectors without inherent sequential memory (unlike Recurrent Neural Networks such as LSTM or Transformer architectures), temporal context must be explicitly encoded through windowed statistics computed over the recent history of each timing source.

1) Rolling Mean $\mu_t(W)$:

Computed as a Finite Impulse Response (FIR) low-pass filter over a window of size W :

$$\mu_t(W) = \frac{1}{W} \sum_{i=0}^{W-1} x_{t-i} \quad (4)$$

This metric suppresses high-frequency symmetric PDV noise to isolate systematic offsets caused by path asymmetry or frequency drift.

Physical Interpretation: μ_t removes transient fluctuations to reveal the underlying static timing error. If $|\mu_{TE}|$ exceeds the normative mask threshold (e.g., $\pm 1,100$ ns per G.8271.1), the source is flagged as compromised regardless of other feature values.

2) Rolling Standard Deviation $\sigma_t(W)$:

Computed as the sample standard deviation over window W :

$$\sigma_t(W) = \sqrt{\frac{1}{W-1} \sum_{i=0}^{W-1} (x_{t-i} - \mu_t)^2} \quad (5)$$

This metric approximates the metrological Time Deviation (TDEV) indicator defined in ITU-T G.8260, measuring the energy of the variable signal component (noise power).

Operational Significance: Critically, during the onset of network congestion bursts, the mean μ_t often remains stable due to averaging inertia, whereas σ_t grows exponentially. This divergence is a key signature that standard BMCA—which monitors only declared attributes and instantaneous offsets—cannot detect. By providing σ_t as an explicit input feature, the classifier learns to identify pre-failure congestion patterns and initiate preventive switching before mask violations occur.

3) Gradient Δ_t :

The *numerical* derivative is computed as the rate of change of the rolling mean:

$$\Delta_t = \frac{\mu_t - \mu_{t-1}}{\Delta t} \quad (6)$$

This metric serves as a direct frequency drift indicator. Upon GNSS signal loss (transition to holdover), the local oscillator begins drifting at a constant rate determined by its inherent frequency offset. While the absolute Time Error (TE) accumulates slowly and may remain within compliant bounds for several minutes, Δ_t exhibits an instantaneous step change at the moment of failure. This property enables detection with minimal latency—significantly faster than threshold-based approaches, which must wait for the integrated error to exceed a configured limit.

4) Multi-scale Windowing Strategy:

To resolve the fundamental trade-off between fault sensitivity and noise robustness, all statistical features are computed simultaneously at two distinct temporal scales:

- Short Window ($W_{short} = 10$ s): Provides high sensitivity to rapid changes and microsecond-level transients, enabling instantaneous reaction to sudden failures (e.g., jamming or switching events) but remaining susceptible to high-frequency noise.
- Long Window ($W_{long} = 60$ s): Facilitates stable trend analysis, offering robustness against transient packet delay variation (PDV) noise but introducing an inherent detection delay.

This dual-scale approach equips the classifier with both a "microscopic" view for immediate anomaly detection and a "macroscopic" view for long-term stability analysis, allowing the decision engine to dynamically weight these perspectives based on the current context.

C. Recursive memory: the *prev_decision* feature

A pivotal methodological innovation is the explicit inclusion of the previous classification output y_{t-1} as an input feature x_{prev} for the current decision step t . This establishes a first-order autoregressive feedback loop within the decision process, formally defined as:

$$X_t = [x_t, \mathcal{J}(y_{t-1})] \quad (7)$$

Where $\mathcal{J}(y_{t-1})$ represents the categorical encoding of the active source at time $t-1$. During the training phase, the Random Forest observes that the ground truth oracle—which accesses hidden state variables and applies conservative switching logic—consistently favors source persistence. Consequently, the classifier learns a conditional probability distribution where maintaining the status quo is statistically privileged:

$$\begin{aligned} P(y_t = A \mid x_t, y_{t-1} = A) &> \\ P(y_t = A \mid x_t, y_{t-1} = B) & \end{aligned} \quad (8)$$

This inequality effectively transforms the static decision boundary into a dynamic *hysteresis* zone, the width of which varies according to the feature-space context. To overcome the learned inertia of x_{prev} and trigger a switch, the alternative source must demonstrate significantly superior quality—a threshold that automatically scales with the severity of the detected anomaly.

In contrast to rigid Wait-to-Restore (T_{WTR}) timers, this learned hysteresis offers four distinct advantages:

1. Adaptive: It scales dynamically with the magnitude of the detected degradation.
2. Data-Driven: It eliminates the need for manual threshold configuration.
3. Site-Agnostic: It ensures consistent behavior across heterogeneous deployment sites without local tuning.
4. Optimally Balanced: The trade-off between stability and responsiveness is derived empirically from the statistical structure of the training data rather than from engineering heuristics.

D. Missing Data Handling

In operational environments, User Datagram Protocol (UDP) packet loss frequently results in intermittent gaps within the PTP telemetry stream. Instead of discarding incomplete time steps—a practice that risks breaking temporal continuity and introducing statistical bias—the system implements an inertial hold mechanism coupled with exponential uncertainty propagation:

$$\hat{x}_t = x_{t-1}; \quad \hat{\sigma}_t = \sigma_{t-1} \cdot \lambda^{\Delta t} \quad (9)$$

Where $\lambda > 1$ represents a configurable penalty factor, and Δt denotes the elapsed time since the last valid measurement update. This approach models the physical reality that system entropy increases during unobserved intervals: as the duration of the monitoring gap extends, the trustworthiness of the cached state degrades. The exponential growth of the uncertainty term ($\hat{\sigma}_t$) ensures that extended data gaps eventually force a source re-evaluation or disqualification, effectively preventing stale telemetry from dominating critical synchronization decisions.

5. Random forest classifier design

A. Architectural justification

The Random Forest (RF) architecture was selected over competing deep learning alternatives—such as Long Short-Term Memory (LSTM) networks, Transformers, or Deep Reinforcement Learning (DRL) agents—based on a rigorous trade-off analysis driven by three critical operational constraints specific to O-RAN (Open Radio Access Network) deployment:

1. **Deterministic Low-Latency Inference:** The synchronization control loop requires inference times strictly under 1 ms to prevent servo instability. While deep neural networks often require hardware acceleration (GPUs/TPUs) to achieve such latency, Random Forests execute as a sequence of conditional if-else statements, which are computationally efficient ($O(\text{depth} \cdot N_{\text{trees}})$) and easily compiled into highly optimized C/C++ code for standard embedded CPUs.
2. **Regulatory Explainability:** As critical infrastructure, telecom synchronization systems are subject to stringent audit requirements. Unlike the "black-box" nature of deep learning, where decision logic is encoded in millions of opaque weight parameters, Random Forests offer inherent interpretability. Through Feature Importance metrics (Mean Decrease in Impurity) and decision path tracking, every classification output can be traced back to specific physical thresholds, satisfying the "right to explanation" in automated decision-making.
3. **Noise Robustness without Overfitting:** Neural networks with high parameter counts are prone to memorizing the stochastic noise patterns of PDV training data (overfitting), leading to poor generalization in the field. In contrast, the ensemble averaging mechanism of Random Forests acts as a stochastic filter, naturally suppressing Gaussian and non-Gaussian noise without

requiring complex regularization techniques like dropout or weight decay [6]:

B. Ensemble mechanism

The classifier is constructed as an ensemble of N uncorrelated decision trees, trained via Bootstrap Aggregation (Bagging) on simulation-generated labeled data. To ensure diversity among the base estimators, two levels of randomization are applied:

- **Row Subsampling (Bagging):** Each tree is trained on a bootstrap sample D_i drawn with replacement from the original dataset D , ensuring that individual trees learn from slightly different data distributions.
- **Feature Subsampling (Random Subspace Method):** At each node split, the algorithm considers only a random subset of features with size $k = \sqrt{M}$, where M is the total feature vector dimensionality. This mechanism decorrelates the trees by preventing strong predictors (e.g., instantaneous TE) from dominating the structure of every tree in the forest.

The final prediction $\hat{y}$ is determined by a majority vote (soft voting) across all T trees, which stabilizes the decision boundary against outliers:

$$\hat{y} = \underset{c}{\operatorname{argmax}} \left(\frac{1}{N} \sum_{i=1}^N P_i(c|x) \right) \quad (10)$$

Where $P_i(c|x)$ is the probability estimate of class c by the i -th tree.

C. Bias-variance decomposition analysis

The theoretical superiority of the Random Forest in this domain can be formalized through the decomposition of the expected generalization error:

$$\operatorname{Err}(f) = \operatorname{Bias}^2 + \operatorname{Variance} +$$

$$\operatorname{IrreducibleNoise} \quad (11)$$

Individual decision trees are "low bias / high variance" learners: they can approximate complex non-linear boundaries (low bias) but are highly sensitive to small perturbations in the training data (high variance). The ensemble averaging mechanism preserves the low bias of the individual trees while aggressively reducing the variance.

Crucially, the variance of the Random Forest average is not simply σ^2/N , but is governed by the correlation ρ between the trees:

$$\operatorname{Var}(RF) = \rho \sigma^2 + \frac{1-\rho}{N} \sigma^2 \quad (12)$$

Where σ^2 is the variance of a single tree and ρ is the average correlation between any pair of trees in the forest [7]. As $N \rightarrow \infty$, the second term vanishes, but the first term remains. This highlights the critical role of feature subsampling ($\sqrt{M}$), which minimizes ρ (decorrelates the trees). By effectively decoupling the estimators, the Random Forest achieves an optimal bias-variance trade-off for the noisy, high-dimensional synchronization feature space. This theoretical property is practically verified by the "stability paradox" results: the classifier produces consistent, reliable decisions even under extreme

PDV conditions where single-model approaches would likely oscillate.

6. Experimental results

A. The stability Paradox

The most significant and counterintuitive finding of this study emerges from the high-PDV scenario (Scenario B), where Gamma-distributed noise ($\alpha = 0.5, \beta = 500$) with AR(1) temporal correlation ($\rho = 0.95$) is injected during the interval $t = 1000 - 2000$ s. This configuration produces heavy-tailed delay spikes reaching $2 - 3 \mu\text{s}$, simulating severe backhaul congestion. Table I presents the comparative results between the standard Best Master Clock Algorithm (BMCA) and the proposed AMSA framework.

Table 1. The stability paradox: switching frequency vs. timing quality.

Metric	BMCA (2 switches)	AMSA (1,318 switches)
Max TE (ns)	1,470.71	892.87
P95 TE (ns)	~980	446.51
Mask Violations	209	0
Strategy	Static persistence	Adaptive micro-switching

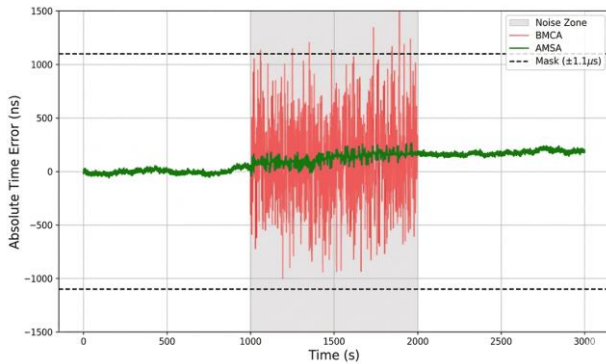


Fig. 2. Time error dynamics under Scenario B. BMCA (red, 2 switches) exhibits chaotic excursions breaching the $\pm 1,100$ ns mask; AMSA (green, 1,318 switches) maintains stability through adaptive micro-switching.

Classical synchronization theory typically posits that optimal systems should minimize source switching, as each transition introduces a risk of phase discontinuity and servo transients. The experimental data presented here contradicts this established principle: the algorithm executing 1,318 switches (averaging one event every 2.7 s during the stress interval) achieves zero mask violations. In stark contrast, the standard algorithm, which performs only 2 switches, records 209 violations and a peak Time Error (TE) dangerously close to the 1,500 ns functional failure threshold.

The resolution of this paradox lies in the fundamental distinction between reactive and proactive switching strategies. The BMCA's 2 switches are purely reactive: the algorithm selects a source and persists with it regardless of measured signal degradation, switching only when declared PTP attributes (such as clockClass) change. However, in this scenario, the grandmaster's attributes

remain unchanged as the congestion is downstream, rendering BMCA blind to the local noise environment.

Conversely, AMSA's 1,318 switches represent a proactive strategy. The Random Forest classifier continuously evaluates the feature vector—particularly the rolling standard deviation σ_{TE} , which captures real-time noise energy—and dynamically alternates between the primary source, alternative sources, and holdover mode. This behavior effectively constructs an optimal composite timing signal. Each micro-switch functions as a non-linear filter that "gates out" PDV-contaminated packet bursts, seamlessly transitioning to a stable holdover or alternative reference during noise spikes, and restoring the primary source only once the transient subsides. This finding redefines stability in modern networks: it is not the absence of change, but the capability to dynamically adapt to stochastic network conditions.

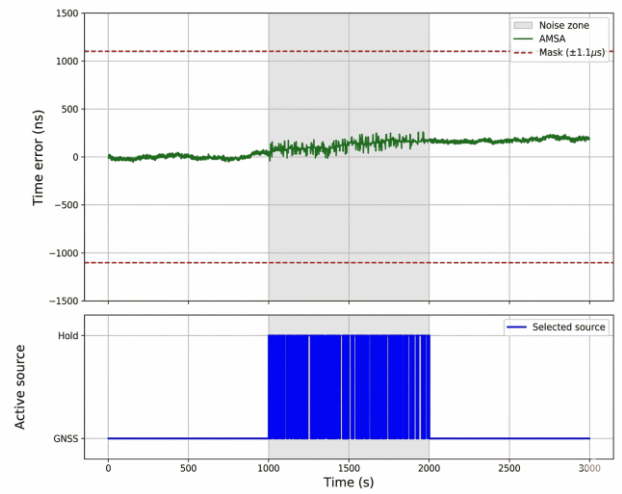


Fig. 3. Dual-panel view of Scenario B: upper panel shows TE remaining within safe bounds throughout; lower panel reveals the dense block of source state transitions during the PDV stress interval ($t=1000-2000$ s), illustrating the micro-switching strategy.

This result fundamentally redefines the concept of operational stability in modern telecommunications: true timing stability is achieved not through static configuration persistence, but through high-frequency dynamic adaptivity. This observation marks a paradigm shift from conventional synchronization management philosophy, suggesting that resilience in stochastic networks requires continuous, proactive state re-evaluation rather than passive adherence to initial settings.

B. Feature importance analysis

To quantify the contribution of each input variable to the classification decision, we conducted a Mean Decrease in Impurity (MDI) analysis—also known as Gini importance—across the entire Random Forest ensemble. This metric aggregates the total reduction in the Gini impurity criterion achieved by splits on a given feature, averaged over all trees in the forest. Table II reports the ranked feature importances computed specifically under adversarial conditions (combining data from Scenarios B and C), where discriminative power is most critical for maintaining synchronization stability.

Table 2. Feature importance ranking (MDI, Adversarial Scenarios).

Rank	Feature	MDI	Physical Meaning
1	σ_{TE} (rolling std)	0.247	Signal instability energy
2	x_{prev} (prev_decision)	0.203	Contextual memory / hysteresis
3	MTIE	0.159	Long-term drift accumulation
4	μ_{TE} (rolling mean)	0.138	Systematic offset indicator
5	TE (instantaneous)	0.112	Noisy single-sample signal
6	PDV α (Gamma shape)	0.087	Network load proxy

The dominance of σ_{TE} and x_{prev} in the feature importance ranking empirically confirms two central hypotheses of this work:

Hypothesis 1 — Instability over accuracy:

The analysis demonstrates that signal instability (σ_{TE}) is a significantly more robust indicator of source quality than absolute Time Error (TE). During Packet Delay Variation (PDV) bursts, the instantaneous TE exhibits rapid, stochastic fluctuations—ranging from 50 ns to 2,000 ns within a single sample window—offering no reliable threshold for rejection. In contrast, the rolling standard deviation reveals a sustained increase in variance that unambiguously signals link degradation. This finding explains the failure mode of the standard BMCA: by monitoring only declared attributes and instantaneous offsets, the legacy algorithm lacks the statistical memory required to distinguish between a transient spike and sustained instability, rendering it blind to congestion-induced quality degradation.

Hypothesis 2 — Memory as the missing component:

The high ranking of x_{prev} (recursive memory) establishes that knowledge of the system state at time $t - 1$ is nearly as informative as the current network observation at time t . This feature empowers the classifier to differentiate between two operationally distinct scenarios that may present identical feature values: a genuine change in conditions warranting a switch, versus stochastic noise hovering around the decision boundary. Furthermore, the relatively low rank of raw, instantaneous TE (5th position) reinforces the conclusion that unprocessed telemetry holds limited classification value compared to derived statistical trends.

Ablation study results

To rigorously quantify the marginal contribution of each architectural component, we conducted a systematic ablation study by removing individual features. The results highlight the critical dependencies of the model:

- Removing σ_{TE} : Causes a critical degradation in $F1$ score from 0.94 to 0.76 ($\Delta = -0.18$). Without noise energy measurement, the classifier loses the ability to distinguish benign PDV bursts from genuine source degradation.

- Removing prev_decision: Results in a severe drop to $F1 = 0.82$ ($\Delta = -0.12$), reintroducing "flapping" instability in boundary zones.
- Removing Multi-scale Windowing: Yields a moderate decline to $F1 = 0.87$ ($\Delta = -0.07$).
- Removing Gradient Δ_t : Lowers performance to $F1 = 0.89$ ($\Delta = -0.05$).
- Removing Z-score Normalization: Has a minor impact ($F1 = 0.91$, $\Delta = -0.03$), confirming the Random Forest's inherent scale invariance, although normalization remains essential for $+500\sigma$ out-of-distribution generalization.

C. Oscillator sensitivity analysis

To quantify the interaction between algorithmic intelligence and hardware quality, all simulation scenarios were repeated across two distinct oscillator classes: TCXO (± 50 ppb frequency stability, representing standard access-layer equipment) and Rubidium (± 0.05 ppb, representing high-cost aggregation-layer or metro nodes).

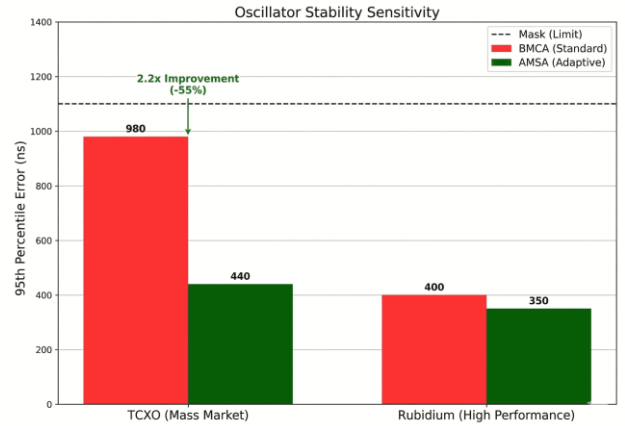


Fig. 4. P95 TE for all four algorithm-oscillator combinations. The critical crossover: AMSA+TCXO outperforms BMCA + Rubidium, proving software intelligence substitutes for hardware investment.

Table 3. Oscillator sensitivity: P95 TE across configurations.

Configuration	P95 TE (ns)	Cost/unit	Relative Performance
BMCA + TCXO	~980	€100	Near-threshold, at-risk
BMCA + Rubidium	~500	€2,000	Adequate margin
AMSA + TCXO	~440	€100	Excellent margin
AMSA + Rubidium	~430	€2,000	Marginal improvement

The underlying mechanism driving this performance is the "Virtual Rubidium" concept. AMSA's polynomial regression model (degree $k = 2 - 3$) actively predicts and compensates for the TCXO's second-order drift during holdover, effectively linearizing the error curve that the standard BMCA allows to grow quadratically. This algorithmic compensation transforms inexpensive TCXO hardware into a virtual atomic reference. In contrast, the BMCA's static holdover logic accumulates drift regardless of oscillator quality because it lacks the predictive model required to compensate for temperature-dependent frequency variations.

D. Explainability in practice

In contrast to "black-box" deep learning architectures, the Random Forest classifier offers inherent interpretability, providing a comprehensive, deterministic audit trail for every synchronization decision. This transparency is achieved by logging the specific feature set contributing to the inference, their relative importance weights (via Mean Decrease in Impurity), and the precise threshold crossings within the dominant decision paths of the ensemble. This capability empowers network operators to address concrete diagnostic queries regarding automated behaviours:

Scenario 1: Justification for Source Switching.

- Query: "Why did the system initiate a switch from GNSS to PTP at time t ?"
- Audit Response: "The rolling noise energy σ_{TE} exceeded 200 ns (3σ threshold) concurrent with a drop in the PDV shape parameter $\hat{\alpha}$ below 0.3, indicating heavy-tailed delay injection. Consequently, the computed Trust Score fell below the critical threshold $T_{crit} = 0.5$, and the ensemble consensus shifted to favor PTP."

Scenario 2: Justification for Stability (Noise Rejection).

- Query: "Why did the system maintain the current source despite a momentary Time Error (TE) spike of 900 ns?"
- Audit Response: "The recursive memory feature x_{prev} identified the current source as 'GNSS' with high confidence ($P = 0.92$). The observed spike was analyzed and determined to lie within the expected σ_{TE} range for the current PDV profile, and no sustained trend was detected in the frequency drift gradient Δ_f . Therefore, the GNSS Trust Score remained above the rejection threshold (0.8)."

This level of traceability satisfies the stringent auditability requirements mandated for critical telecommunications infrastructure, enabling post-incident forensic analysis that is unattainable with the deterministic logic of standard BMCA.

Performance and Granular Analysis.

Per-class classification metrics confirm balanced accuracy across all operational modes, yielding the following F1 scores:

$$\begin{aligned} F1_{GNSS} &= 0.96, & F1_{PTP} &= 0.93, \\ F1_{Holdover} &= 0.92 \end{aligned} \quad (13)$$

Resulting in a macro-average of $F1 = 0.94$.

A concrete example of this granular explainability is observed at $t = 1,047$ s: the system executed a switch from GNSS to PTP driven by specific conditions— σ_{TE} exceeded 3.2σ (against a 2.8σ dynamic threshold), $\hat{\alpha}$ dropped to 0.31 (indicating a heavy-tailed distribution), and the GNSS Trust Score degraded to 0.43 (below $T_{crit} = 0.5$). Simultaneously, 87% of the constituent decision trees voted for the PTP class. This detailed log data enables precise root-cause analysis of network anomalies.

7. Conclusion

This paper presented a feature engineering and classification methodology for explainable time source arbitration in 5G networks. Three principal results were established:

First, multi-scale rolling statistics ($\mu_t, \sigma_t, \Delta_t$) with Z-score normalization transform noisy PTP telemetry into semantically rich features that enable ML-based classification where threshold-based approaches fail. The dual-scale windowing (10 s / 60 s) provides both transient detection sensitivity and trend analysis robustness.

Second, the recursive `prev_decision` feature implements data-driven adaptive hysteresis that eliminates flapping without manual tuning—a fundamental advance over static wait-to-restore timers that require per-site configuration and periodic recalibration.

Third, the stability paradox demonstrates that adaptive micro-switching (1,318 transitions with zero violations) is categorically superior to static persistence (2 transitions with 209 violations) under adversarial conditions, challenging the conventional assumption that minimal switching implies optimal stability.

Feature importance analysis confirmed that σ_{TE} and x_{prev} dominate classification in adversarial scenarios, validating the hypothesis that contextual memory is the missing component in conventional synchronization management. Oscillator sensitivity analysis demonstrated that AMSA + TCXO outperforms BMCA + Rubidium, proving that algorithmic intelligence compensates for hardware limitations.

Future work will extend the explainability framework to real-time operator dashboards with interactive decision visualization, investigate attention-based sequence models as alternatives to rolling statistics for temporal feature extraction, and explore federated feature engineering across multi-operator timing domains to enable cross-network learning without compromising data sovereignty.

8. Future work

Future research will focus on extending the explainability framework to real-time operator dashboards featuring interactive decision visualization. Additionally, we will investigate attention-based sequence models (e.g., Transformers) as potential alternatives to fixed rolling statistics for temporal feature extraction and explore federated feature engineering across multi-operator timing domains to enable cross-network learning without compromising data sovereignty.

References

- [1] G. P. Sharma et al., "Toward Deterministic Communications in 6G Networks: State of the Art, Open Challenges and the Way Forward," IEEE Access, vol. 11, pp. 106898–106923, 2023.
- [2] IEEE, "IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems," IEEE Std 1588-2019, Jun. 2020.
- [3] M. Lévesque and D. Tipper, "A Survey of Clock Synchronization Over Packet-Switched Networks," IEEE Commun. Surveys Tuts., vol. 18, no. 4, pp. 2926–2947, 2016.
- [4] F. Girela-López et al., "IEEE 1588 High Accuracy Default Profile: Applications and Challenges," IEEE Access, vol. 8, pp. 45211–45220, 2020.

-
- [5] D. Chandramouli, P. Andrés-Maldonado, and T. Kolding, "Evolution of Timing Services From 5G-A Toward 6G," *IEEE Access*, vol. 11, pp. 35150–35157, 2023.
- [6] C. Zhang et al., "Three-Way Selection Random Forest Optimization Model for Anomaly Traffic Detection," *Electronics*, vol. 12, no. 8, p. 1788, 2023.
- [7] G. Giorgi, "An Event-Based Kalman Filter for Clock Synchronization," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 2, pp. 449–457, 2015.
- [8] Y.-D. Seo et al., "Improved Time-Synchronization Algorithm Based on Direct Compensation of Disturbance Effects," *Sensors*, vol. 19, no. 16, p. 3499, 2019.
- [9] Z. Wang et al., "Time Synchronization for 5G and TSN Integrated Networking," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 9, pp. 2969–2980, 2025.
- [10] A. Nasrallah et al., "Ultra-Low Latency (ULL) Networks: The IEEE TSN and IETF DetNet Standards," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 88–145, 2019.
- [11] T. Iliev and I. Penkov, "AI-Driven Trust Score-Based Synchronization Control for Robust 5G and Beyond Networks," in *Proc. ICAI 2025*, pp. 504–509.
- [12] P. Assimakopoulos et al., "A Converged Evolved Ethernet Fronthaul for the 5G Era," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2528–2537, 2018.

Assessment of Digital Infrastructure Maturity in Bulgarian Scholarly Libraries: Empirical Study of Automation and AI Adoption

Viktoriya Ivanova

University Library, University of Ruse “Angel Kanchev”, Ruse, Bulgaria

*E-mail of corresponding author: vivanova@uni-ruse.bg

Abstract: Scholarly libraries play central role in providing services for introduction and management of scientific data, including infrastructures for institutional repositories and metadata support. In the context of complex digital research activities, scholarly libraries support access, preservation and distribution of digital resources. AI supported services give new perspectives to information sciences and widen opportunities for innovations and cooperation. Purpose of the article is to assess the digital infrastructure maturity in Bulgarian scholarly libraries and to perform empirical study of automation and AI adoption. For achieving this goal, a survey is carried out among Bulgarian scholarly libraries. It outlines the infrastructural capacity and the basic parameters, necessary for the introduction of a measurement model. The Curation Lifestyle Model is adapted to perform quantity measurement of parameters and calculate a Digital maturity index (DIMI). After that, Pearson chi-square test is performed to check the correlation between the size of the university and the DIMI. Finally, Spearman correlation formula is used to establish a relation between DIMI and AI adoption.

Received 25 March 2026

Accepted 03 June 2026

Published 28 August 2026

Keywords: Digital maturity index, Libraries, Chi-square test, Spearman correlation.

1. Introduction

Scholarly libraries are the main repositories for scientific data and technological evolution turned them into innovative digital hubs. They develop virtual collections in online environments and integrate intelligent systems in every aspect of their operation. Strategic plans and management policies of research libraries include digital transformation, innovation of services, and remodeling to meet the demands of higher education institutions [1].

The rising importance of technology and data sharing has pushed research libraries to develop digital services in order to serve their communities best [2]. They constantly upgrade their technological infrastructures, enrich online resources, improve cataloguing services, and, at the same time provide user-friendly interfaces and continuous access to information. An integral prerequisite for the development of contemporary scholarly libraries is IT infrastructures, which support digital services and digital management. Next steps are automation of cataloguing, user analysis, discovery services, and statistics. Inevitable parts of the well-developed scholarly library are also electronic references, research data support, digitization services, etc. Open science frameworks and affirmed legislation policies regulate standards for metadata and support interoperability – Open AIRE, FAIR, OAIS [3]. For a better understanding of data and its proper use and interpretations, metadata standards are implied.

Depending on the structure, value, contents and interchange models, there are different types of metadata standards – MARC, MODS, Dublin Core, etc. [4].

Currently, science is in a position where digital data, automation of services, and AI-support systems determine its level of excellence. More recently, artificial intelligence technologies have begun to influence library services by enabling automated metadata enrichment, intelligent discovery systems, and advanced analytics [5].

In the near future automated systems will turn into expert systems, reference services will transform into Chatbots and voice assistants. Traditional classification schemes will be converted to text classification and semantic indexing [6].

When digital resources are managed properly, the institutions widen their opportunities for productivity and community cooperation. A variety of digital curation techniques are offered in scientific literature for better storage and maintenance of digital collections [7]. In this context, scholarly libraries are starting to implement digital management policies and strategies, which support them in the archiving and curation of educational and scientific data [8].

The Curation Lifestyle Model, developed by the Digital Curation Center, provides an overview of the stages of data preservation. The model serves as a basis for creating a framework of requirements and technological infrastructures for maintaining, preserving, and adding value to digital research data throughout its lifecycle [9]. Evaluation of the implementation of DCC Curation Lifestyle Model is possible with the use of the Digital Maturity Model (DMM). It performs a quantity measurement of parameters to determine the technological



© The Author(s). Published by CIEES Association Press. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 Generic License.

infrastructure, identifies possible improvements and underlines development opportunities [10].

Bulgarian scholarly libraries, as part of the European and World scientific area, are expected to develop with high intensity and to introduce modern digital infrastructures and services. These activities ensure the affirmation of Bulgarian scientists in the international scientific community and create a wide range of cooperation opportunities, and innovation, and knowledge exchange.

The purpose of this paper is to assess the Digital Infrastructure Maturity Index (DIMI) in Bulgarian scholarly libraries and to perform an empirical study in order to evaluate if there is a correlation with size of the institution and AI adoption. Three research questions are defined:

1. Assessment of DIMI in Bulgarian scholarly libraries.
2. Test of the relationship between university size and DIMI.
3. Calculation of the correlation between DIMI and AI adoption.

2. Materials and Methods

The defined research questions and the selected methods for their investigation have been listed in Table 1, below. The first research question aims to reveal the current state of digital and technological development of Bulgarian scholarly libraries through measuring a Digital Infrastructure Maturity Index. For this purpose, the DCC Curation Lifestyle Model is adapted to the Digital Maturity Model and quantity analyses of parameters is performed. After that Chi-square test investigates if there is a relation between the size of the university and the DIMI. Finally, with a look to-wards the near future, a calculation with the Spearman correlation formula is performed in order to establish if there is a correlation between the DIMI and the AI adoption in Bulgarian scholarly libraries.

Table 1. Research questions and methods for their investigation.

Research Question	Method
1. Measurement of the DIMI in Bulgarian scholarly libraries	DIMI
2. Relationship between size of the university and Digital maturity index (DIMI)	Chi-square (χ^2) test
3. Correlation between adoption of AI and Digital infrastructure maturity (DIMI)	Spearman correlation

2.1. Digital maturity index (DIMI) in Bulgarian scholarly libraries

For determining the DIMI of Bulgarian scholarly libraries, an evaluation framework is adapted, based on the DCC Curation Lifestyle Model and the principles of the Digital Maturity Models.

The Digital Maturity Model is an instrument for establishing the level of digital maturity in organizations. It measures the extent to which a company implements digital technologies and processes through various parameters (strategy, technology, service, etc.) Each parameter is evaluated on the basis of a maturity scale (for

example, between 1 and 5). The total sum shows the current level of digitalization [11].

$$\text{DIMI} = \sum_{i=1}^N \text{Parameter}_i \quad (1)$$

The DIMI is calculated with the use of parameters based on the DCC Curation Lifestyle Model.

The DCC Curation Lifestyle Model provides overview of the stages of data preservation. It serves as a basis for creating a framework of requirements and techno-logical infrastructures. There are 8 sequential actions:

1. Conceptualize – initiate and plan the creation of data (gathering and preservation).

2. Create and receive – create data, including administrative, descriptive, structural, and technical metadata. Receive data from various sources in accordance with policies, and put the correct metadata.

3. Appraise and select – evaluate data and select for long-term preservation.

4. Ingest – transfer data to an archive, repository or other custodian.

5. Preservation action – provide long-term preservation. Data cleaning, validation, assigning preservation metadata, assigning representation information, and providing acceptable data structures and formats.

6. Store – keep the data, complying with relevant standards.

7. Access, Use, and Reuse – ensure that data is constantly accessible. In some cases, authentication and access control are applied.

8. Transform – create new data from the original (different format or subset) [12].

A questionnaire is designed for Bulgarian scholarly libraries with the title: “Digital Library Information Systems and Services”. It includes 14 questions, united under 6 sections: General university information; Core library information systems; Digital services and automation; Interoperability and standards; Intelligent and advanced services; International cooperation.

For the purposes of this research, six questions are selected and aligned with the 8 sequential actions of the DCC Curation Lifestyle Model:

1. Does your institution use an institutional repository or digital collection?

2. Which digital library services are currently automated?

3. Does your library provide any of the following digital services?

4. Which metadata standards are used?

5. Is the library aligned with any of the following frameworks or guidelines?

6. Is the library currently using or plan to use intelligent / AI-supported services?

After defining the correspondence between the DCC Curation Lifestyle Model and the selected survey questions, an evaluation framework is designed by attributing scoring points for each answer. After the aggregation of all score points, the DIMI for each Bulgarian scholarly library is calculated.

Table 2 illustrates the correspondence between the DCC Curation Lifestyle Model, the survey questions and the score attributed to each answer.

Table 2. Framework for evaluation of the DIMI.

Curation Lifestyle Model →	Survey Question →	Score
Conceptualize	Institutional Repository	
Create and receive data	Yes	Yes – 1
	No	No – 0
Curation Lifestyle Model →	Survey Question →	Score
	Automation of library services	
Appraise and select -	Cataloguing and metadata management	(0-1) services – 0
Evaluate data and select	Circulation and user management	(2-3) services – 1
for preservation	Digital content ingests and workflows	(4-5) services – 2
	Discovery and search services	
	Usage statistics and analytics	
	None / limited automation	
Curation Lifestyle Model →	Survey Question →	Score
	Digital services provided	
Ingest - transfer data to	Remote access to electronic resources	(0-1) services – 0
an archive, repository,	Digital reference services	(2-3) services – 1
data center	Research data management support	(4-5) services – 2
	Open access publishing support	
	Digitization services	
Curation Lifestyle Model →	Survey Question →	Score
	Metadata standards	
Preservation action -	Dublin Core	Local only – 0
data cleaning, validation,	MARC21	One foreign standard – 1
acceptable file formats	MODS	Two and more foreign
	Local	standards – 2
Curation Lifestyle Model →	Survey Question →	Score
	Alignment with frameworks	
Store - coherence	Open AIRE Guidelines	None – 0
with standards	FAIR principles	At least one – 1
	OAIS model	
	None	
Curation Lifestyle Model →	Survey Question →	Score
	AI-supported services	
Access, use and reuse	Yes – currently implemented	Yes – 2
Transform	Planned within 1–2 years	Planned – 1
	No current plans	No – 0
Total score →	Maturity Index Scale	
Total score - 11	Low maturity – 0-3	
	Moderate maturity – 4-6	
	Advanced maturity – 7-11	

2.2 Relationship between the size of the university and DIMI

A Chi-square index (χ^2) is a statistical indicator, which is used to establish whether there is a significant correlation or difference between categorical variables. It compares the observed values with the expected values, submitted in a table. If the value of χ^2 is big enough and the p-value (the calculated result) is lower from the determined level, it is concluded that there is a statistical correlation between the variables [13].

A Chi-square test of interdependence was carried out to check if there is a relationship between the size of the university and the DIMI of scholarly libraries. For this purpose, two Hypothesis were determined:

H₀: There is no interdependence between the size of the university and DIMI.

H₁: There is a relationship between the size of the university and DIMI.

The Pearson Chi-square test follows the steps:

- Creation of a table with observed and expected frequencies.

- Calculation of the chi-square value with Excel.
- Comparing the value to the critical value.
- Deciding whether to reject the null hypothesis.

2.3 Correlation between adoption of AI and DIMI formatting

The relationship between the Digital infrastructure maturity and AI adoption in academic libraries was examined with the Spearman correlation.

Spearman correlation is a statistical method that measures the strength and direction of the relation between two variables, while the data has been arranged in ranks. It is used with non-normally distributed data or ordinal data and shows if there is a monotonic relationship (the two variables change in the same direction).

The Spearman correlation is calculated not with the raw (initial) data, but instead with ranked values. The ranked values are arranged in order from lowest to highest, and a number is assigned to each value, based on its new position

Spearman correlation (ρ or r_s) formula:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \quad (2)$$

where: d_i is the difference between the ranks of the two variables for each observation; n is the number of observations. The coefficients range from -1 to 1, where -1 is a perfect negative correlation, 0 indicates no correlation, and 1 is a perfect positive correlation [14].

3. Results and Discussion

3.1. Digital maturity index (DIMI) in Bulgarian scholarly libraries.

The questionnaire “Digital Library Information Systems and Services” was prepared and distributed among Bulgarian scholarly libraries in the period January–March 2026.

In 2026, the total number of scholarly libraries listed in the Register of Higher Education Institutions in Bulgaria is 51. The survey was distributed to all libraries. Participation in the survey was voluntary. Twenty-nine scholarly libraries gave their answer, which corresponds to approximately 56.9 %. The study therefore, represents a partial census of the population of scholarly libraries and provides a substantial representation.

The main purpose of the survey is to gather accurate information regarding the digital library information systems, infrastructures, and services, which are implemented in Bulgarian scholarly libraries.

The survey includes 14 questions, united under six sections: General university information; Core library information systems; Digital services and automation; Interoperability and standards; Intelligent and advanced services; International cooperation. For the purposes of the research questions, defined in this paper, six questions have been selected. Tables 3, 4, 5, 6, 7, and 8 present the answers given by the Bulgarian scholarly libraries.

From the analysis of the answers for Question 1: Does your institution use an institutional repository or digital collection, presented in Table 3, it can be summarized that 62.1 % from the scholarly libraries possess an institutional repository or digital collection. Some libraries operate an institutional repository – 20.69 %, others manage digital collections for cultural heritage – 17.24 %. The libraries, which have both – a repository and digital collections, are 24.14 %.

Table 3. Does your institution use an institutional repository or digital collection? (Question 1).

Answer	N scholarly libraries	%
Yes – both	7	24.11
Yes – digital collections for cultural heritage	5	17.24
Yes – institutional repository	6	20.69
No	11	37.93

Table 4 illustrates the answers to Question 2: Which digital library services are currently automated. Circulation and user management services have the highest percent of automation – 89.66 %, followed by discovery and search services – 82.75 % and cataloguing

and metadata management – 79.31 %. Relatively low level of automation is observed in the digital content ingest and workflows, and the statistics – each with 65.52 %. There is one scholarly library that has not automated its processes yet.

Table 4. Which digital library services are currently automated? (Question 2).

Answer	N scholarly libraries	%
Cataloguing and metadata management	23	79.31
Circulation and user management	26	89.66
Digital content ingests and workflows	19	65.52
Discovery and search services	24	82.76
Usage statistics and analytics	19	65.52
None / limited automation	1	3.45

Table 5 represents the distribution of digital services among Bulgarian scholarly libraries. The highest percent is 89.66 %, and it is for providing remote access to electronic resources, followed by digital reference services – 68.97 %, and digitization services – 58.62 %. The lowest percent is for the research data management support – 37.93 %.

Table 5. Does your library provide any of the following digital services? (Question 3).

Answer	N scholarly libraries	%
Remote access to electronic resources	26	89.66
Digital reference services	20	68.97
Research data management support	11	37.93
Open access publishing support	13	44.83
Remote access to electronic resources	17	58.62

Table 6 shows the answers to Question 4: Which metadata standards are used. Local standards are implemented in more than half of the scholarly libraries – 57.69 %. The rest are using either MARC 21 – 38.46 %, or Dublin Core – 23.08 %. None of the Bulgarian scholarly libraries has implemented MODS.

Table 6. Which metadata standards are used? (Question 4).

Answer	N scholarly libraries	%
Dublin Core	6	23.08
ARC21	10	38.46
MODS	0	0
Local	15	57.69

Table 7 shows to what extent Bulgarian scholarly libraries are aligned with the international frameworks for managing digital data – Question 5. The answers show that 79.31% are not using any international standards. Open AIRE Guidelines are used by 13.79 % and FAIR principles are implemented in 6.90 % from the libraries. The OAIS model has not been introduced in Bulgarian scholarly libraries yet.

Table 7. Is the library aligned with any of the following frameworks or guidelines? (Question 5).

Answer	N scholarly libraries	%
Open AIRE Guidelines	4	13.79
FAIR principles	2	6.90
OAIS model	0	0
None	23	79.31

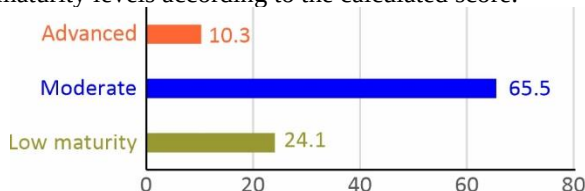
Question 6 explores the innovative services offered by scholarly libraries: Is the library currently using or plan to use intelligent / AI-supported services. Table 8 represents the answers to Question 6 – 53.57 % have either planned or already implemented AI services. The rest 46.43 % have no plans yet of introducing intelligent technologies.

Table 8. Is the library currently using or plan to use intelligent / AI-supported services? (Question 6).

Answer	N scholarly libraries	%
Currently implemented	1	3.57
Planned within 1–2 years	14	50.00
No current plans	13	46.43

The questions and answers from the survey are aggregated according to the framework in Table 2. Six indicators are formed, related to repository infrastructure, digital services, metadata standards, interoperability frameworks, and AI adoption. Each indicator was assigned a score, and the total sum was then aggregated to form a DIMI.

The index was ranked in three levels – Low, Moderate, and Advanced levels of DIMI in Bulgarian scholarly libraries. Fig. 1 illustrates the distribution of the three maturity levels according to the calculated score.

**Fig. 1.** DIMI of Bulgarian scholarly libraries.

In Bulgaria, currently 10.3 % of the scholarly libraries have reached an advanced level of DIMI. The majority – 65.5 % have reached a moderate level of digital transformation. Almost one quarter – 24.1 % are still in the position of low infrastructure maturity level.

3.2 Relationship between the size of the university and DIMI

The second research question - Test of the relationship between university size and DIMI, was investigated with the use of the Chi – square test.

Table 9 represents the distribution of universities, according to the number of their students. It is observed that the greatest number are the universities with less than 10 000 students. Universities with students between 10 000 and 20 000 are 13.79 %. Only 6.90% are the scholarly libraries in universities with a number of students over 20 000.

Table 9. Size of the University (Number of students).

Answer	N scholarly libraries	%
Below 10 000	23	79.31
10 000 – 20 000	4	13.79
Over 20 000	2	6.90

After defining the size of the universities, the data was supplied in Table 10, where the DIMI was already calculated and added. Low DIMI have seven scholarly libraries; Moderate DIMI have 19 libraries, and Advanced DIMI - 3 libraries.

Chi-square formula was applied. Two hypotheses were determined:

H0: There is no interdependence between the size of university and Digital Infra-structure Maturity Index (DIMI)

H1: There is a relationship between the size of university and DIMI.

The degrees of freedom were automatically calculated with a value 4.

The critical value (alpha is 0.05).

The result from the Chi-square test is 47.93. H_0 is rejected.

H_1 is accepted - There is a relationship between the size of university and DIMI of scholarly libraries in Bulgarian universities.

Table 10. Chi- square test algorithm.

Observed Frequencies				
University Size	Low DIMI	Moderate DIMI	Advanced DIMI	Row Total
<10k	7	14	2	23
10k–20k	0	4	0	4
>20k	0	1	1	2
Column Total	7	19	3	29
Chi-square statistic	47.93103448			
Degrees of freedom	4			
Critical value (alpha=0.05)	9.487729037			
Decision	Reject H_0 (relationship exists)			

3.3 Correlation between adoption of AI and DIMI

The Spearman correlation formula was used for solving the third research question, defined in the paper: Calculation of the correlation between DIMI and AI adoption.

Data about the DIMI of Bulgarian scholarly libraries: Low DIMI have a total of 7 scholarly libraries; Moderate DIMI have 19 libraries and Advanced DIMI - 3 libraries.

For the formula necessary results are also answers from the survey question: Is the library currently using or plan to use intelligent / AI-supported services. The number of libraries, which have already implemented AI services is one, and those, which are planning to introduce such technologies are 14. No intention yet of using intelligent/AI supported services have 13 scholarly libraries.

Table 11 displays the results from the Spearman correlation calculation – 0.51. According to the given scale, it is established that there is a Moderate correlation between the DIMI and the adoption of AI-supported services.

Table 11. Spearman Correlation (DIMI vs AI).

Spearman rho:	0.51
Interpretation Guide	
0 – 0.19	Very weak
0.20 – 0.39	Weak
0.40 – 0.59	Moderate
0.60 – 0.79	Strong
0.80 – 1.00	Very strong

3.4 Discussion

The paper investigates three research questions regarding the digital transformation in Bulgarian scholarly libraries. The results show that the majority of libraries have reached a moderate level of Digital technological infrastructure maturity. They operate institutional repositories and digital collections, and most of the processes are automated. Access, preservation and distribution of digital data are managed in a proper way, according to either national or international rules. Bulgarian scholarly libraries assist researchers in their scientific work and support collaboration in international academic communities. Intelligent/AI instruments are acknowledged and their adoption is planned for the near future. These results show good understanding of the future of the scholarly library – continuous development and service variety.

Additional support from authorities and institutions can be directed towards the full automation of library services. The basis for development of digital and AI services is technological infrastructure and complete automation.

Results also indicate that the level of research services in scholarly libraries can be improved by hiring additional experts or introducing training. The implementation of international metadata standards and regulations is

mandatory for interoperability of information, and for expanding of the cooperation.

The research established relationship between the size of the university and the DIMI. The higher the number of students in one university, the higher the demand for contemporary data in multiple formats. More funding and expertise are transferred to larger universities, which leads to advanced technological structures, automation of workflow, and digitalization of services. Important for conducting research and developing innovations is also smaller specialized scholarly libraries. Cutting-edge science requires funding and new investigations; therefore, smaller, but boutique libraries should also participate in the technological transformation in Bulgarian universities.

The Spearman correlation revealed a Moderate level of interdependence between DIMI and AI adoption. The well-developed digital infrastructure in scholarly libraries supports the introduction of AI-supported services. Although intelligent services require certain technological parameters, infrastructure is not the only factor. Introduction of AI technologies in university operations are also connected with the decision of the management and an implementation plan, aligned with international laws and regulations.

4. Conclusion

This study measures the DIMI in Bulgarian scholarly libraries and evaluates its correlation with size of the institution and AI adoption. The results can be summarized as follows:

1. Digital Infrastructure Maturity Index in Bulgarian scholarly libraries. The majority – 65.5 % have moderate level of digital infrastructure maturity; almost one quarter – 24.1 % have low level and 10.3 % show advanced level of DIMI.

2. Test of the relationship between university size and DIMI. Chi – square test is performed and it reveals that there is a relationship between the size of the institution and DIMI.

3. Calculation of the correlation between DIMI and AI adoption. The Spearman correlation shows that there is a moderate correlation between the Digital Infrastructure Maturity Index and the adoption of AI-supported services.

The internationalization of science requires alignment of the infrastructures for data storage and exchange. Technological infrastructures, information systems, and metadata standards should be interoperable.

The survey “Digital Library Information Systems and Services”, distributed among Bulgarian scholarly libraries reveals very good technological capacity, well developed digital services and relatively modern library information systems. A step forward internationalization would be the development of shared infrastructures and the building of compatible systems on national and European level. The results after the calculation of the Digital Infrastructure Maturity Index also show that Bulgarian scholarly libraries can offer a stable infrastructural basis for further development and international data exchange.

Bigger and multidisciplinary institutions are more competitive, possess larger amount of funding and invest more in innovations. They should play the role of leaders in the adoption of international practices and policies. At

the same time, AI instruments allow smaller and not so technologically advanced scholarly libraries to maintain good position and modernize their services.

Further research about the Digital technological infrastructure index among European scholarly libraries will expand the knowledge and provide data for planning future recommendations and actions.

References

- [1] Oyedokun (2024). Navigating the dynamics of present-day academic libraries: An in-depth analysis of strategies, challenges, and emerging trends, *IFLA Journal*. DOI: 10.1177/03400352241291907.
- [2] Tenopir, C., Sandusky, R.J., Allard, S., and Birch, B. (2014). Research Data Management Services in Academic Research Libraries and Perceptions of Librarians. *Library & Information Science Research*, 36, 84-90. <https://doi.org/10.1016/j.lisr.2013.11.003>
- [3] Patel, C. (2025). Recent Advancements in Management and Academic Library Sciences: *The Digital Evolution of Scholarly Libraries*. 163–172. <https://doi.org/10.62823/mgm/9789349468160/13>
- [4] Mosha, N.F.; Ngulube, P. Metadata (2023). Standard for Continuous Preservation, Discovery, and Reuse of Research Data in Repositories by Higher Education Institutions: A Systematic Review, *Information*, 14, 427. <https://doi.org/10.3390/info14080427>
- [5] Cox, A.M., Pinfield, S. and Rutter, S. (2019). The intelligent library: Thought leaders' views on the likely impact of artificial intelligence on academic libraries. *Library Hi Tech*, 37 (3). pp. 418-435. ISSN: 0737-8831 <https://doi.org/10.1108/LHT-08-2018-0105>
- [6] Ali M.Y., Naeem S.B., Bhatti R., Richardson J. (2024). Artificial intelligence application in university libraries of Pakistan: SWOT analysis and implications. *Global Knowledge, Memory and Communication*, Vol. 73 No. 1-2 pp. 219–234, <https://doi.org/10.1108/GKMC-12-2021-0203>.
- [7] Feng Y. and Richards L. (2018). A review of digital curation professional competencies: theory and current practices. *Records Manag J201*, 8; 28(1): 62–78. DOI:10.1108/RMJ-09-2016-0022.
- [8] Hudson-Vitale C., Carlson J.R., Hadley H., et al. (2021). Introduction to the special JeSLIB issue on data curation in practice. *J EScience Librarianship* 10(3): 13.
- [9] Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation* 3(1). Retrieved October 18, from <http://www.ijdc.net/ijdc/article/view/69/69>.
- [10] Gökalp, E. and Martinez, V. (2021). Digital transformation maturity assessment: development of the digital transformation capability maturity model. *International Journal of Production Research*. <https://doi.org/10.1080/00207543.2021.1991020>.
- [11] Minh, H. and Pham, H. (2022). Comprehensive review of Digital Maturity Model and Proposal for a Continuous digital transformation process with Digital Maturity Model integration. *Sistemas & Gestão*. DOI: 10.22937/IJCSNS.2022.22.1.97
- [12] Higgins, S. (2009). DCC DIFFUSE Standards Frameworks: A Standards Path through the Curation Lifecycle. *International Journal of Digital Curation*. DOI: 10.2218/ijdc.
- [13] Tenopir, C., Dane H., Suzie A., Mike F., Ben B., Lynn B., Robert S., Madison L., and Andrew L. (2015). Research Data Services in Academic Libraries: Data Intensive Roles for the Future? *Journal of eScience Librarianship*, 4(2): e1085. <http://dx.doi.org/10.7191/jeslib.2015.1085>.
- [14] Cox, A. (2023). How artificial intelligence might change academic library work: Applying the competencies literature and the theory of the professions. *Journal of the Association for Information Science and Technology*, 74(3), 367–380. <https://doi.org/10.1002/asi.24635>.

Analysis and Prediction of Life Expectancy at the Global Level and in Selected Countries

Dimitrios Kazolis ^{1,*}, Ioanna Kranioti ¹ and Theodosia Kazoli ²

¹ Department of Physics, Democritus University of Thrace, Kavala, Greece; okrani@physics.duth.gr

² Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece; ics26195@uom.edu.gr

*E-mail of corresponding author: dkazolis@physics.duth.gr

Abstract: The relationship between life expectancy and healthcare costs has been the subject of numerous studies by various organizations. This is because it serves as a key tool for evaluating the effectiveness of the healthcare sector, as it affects both the national budget and the quality of life and life expectancy of ordinary citizens. The purpose of this study is to contribute to the advancement of this research by incorporating additional parameters and applying combinations of data analysis techniques. Initially, the differences between selected and representative countries were presented, but basically this analysis focused on predicting life expectancy using a set of economic, health, and environmental indicators. To achieve this goal, two data sets from the World Bank's open data platform were used for the period 2000–2023. The study involved handling of missing values, performing correlation analysis, applying a linear regression model, conducting principal component analysis, and creating a regression model using the random forest method. The process was implemented using the Python programming language. Finally, the summary results were presented and recommendations for future research were formulated. This comprehensive effort is intended to serve as an advisory tool for relevant decision-making and provides the basis for further analysis of the topic.

Received 14 April 2026

Accepted 15 May 2026

Published 1 September 2026

Keywords: Life Expectancy; Health Expenditure; k-Nearest Neighbors; Correlation Analysis; Regression Model; Principal Component Analysis; Random Forest.

1. Introduction

Various projects have been conducted on life expectancy and the factors that influence it [1,2]. This paper seeks to delve deeper into this important topic by using additional data and combining different methodologies. Drawing on a variety of databases from the World Bank [3], an effort is made to identify correlations between these databases, even if they do not necessarily reflect a causal relationship. Thus, databases used that include life expectancy, per capita health expenditure, GDP per capita, the number of hospital beds per 1,000 people, liters of pure alcohol per person aged 15 and older, smoking prevalence including all types of tobacco, and infant mortality for children up to five years of age per 100 births.

In addition, data sets related to education have been collected, such as the percentage of enrollments in higher education relative to the population of the corresponding "standard" age group. Additionally, a separate dataset is provided showing the average number of years of education completed by adults (25+), which is considered to provide greater accuracy, as it further filters out noise

from the previous list. Another dataset is also provided that focuses more on environmental factors, as it examines air pollution, specifically the average concentration of PM2.5 particulate matter in the air in micrograms per cubic meter.

Moreover, records have been included regarding the percentage of the population that has access to drinking water, the percentage of the labor force that is unemployed but actively seeking work, and the percentages of the adult population that is obese, i.e., has a body mass index (BMI) of 30 or higher.

The purpose of this study, therefore, is to examine all the factors that may influence life expectancy. This is achieved through various procedures, as described in the following chapter.

2. Materials and Methods

As mentioned above, As mentioned above, the data were retrieved from the World Bank's open data platform. Initially, they were processed to clean and standardize them. The main challenge was filling in the missing values. The columns with relatively small percentages of missing data (such as 'Obesity', 'Pure Alcohol', 'PM2.5', and 'Education') were imputed with the mean or median value. However, for 'Smoking prevalence %' with over 63% missing data, more advanced imputation techniques were considered. This task was accomplished using the k-Nearest Neighbors (KNN) method [4], which, however, will not be analyzed for reasons of brevity, and thus, the



© The Author(s). Published by CIEES Association Press. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 Generic License.

missing values were imputed by considering patterns in the remaining features.

2.1. Visualization of data

This study examines the correlation not only at the global level but also analyzes data for 22 selected countries. These are: Belgium, Brazil, Bulgaria, Canada, Cyprus, Germany, France, Greece, Italy, Japan, Luxembourg, Norway, Poland, Romania, Portugal, South Africa, Sweden, Switzerland, Spain, Turkey, the United States, and the United Kingdom. Fig. 1 illustrates the relationship between life expectancy and the healthcare costs per citizen.

This graph can effectively be obtained by removing the time component from the two corresponding time series. The Python program was used in the Thonny environment to create this graph. The results shown, are very important for drawing initial conclusions. In general, both variables revile an upward trend over time, with a slight apparent decline during the COVID-19 pandemic, which appears to have had the greatest impact in 2021, the year in which the effects of the global pandemic became apparent.

The general trend suggests that countries with higher expenditures may have better prospects for a longer life expectancy. It can be noted that the US has the highest expenditures; nevertheless, its life expectancy is significantly lower compared to other countries. Japan, on the other hand, achieves the highest life expectancy among the selected countries with very low expenditures, a fact that may be attributed to various socio-economic factors. South Africa appears to follow the pattern of low spending and low life expectancy and has the lowest figures.

Furthermore, Greece has a relatively high life expectancy relative to its expenditures. As for Turkey, Bulgaria, and the global trend, the low life expectancy is justified by the correspondingly low expenditures. Norway and Switzerland have high life expectancy, but also high expenditures.

In addition, the following chart illustrates the variables of life expectancy and healthcare costs for the selected countries, as well as the global trend. These figures are for the year 2023. Fig. 2 better illustrates the strong correlation between life expectancy and health expenditure per capita.

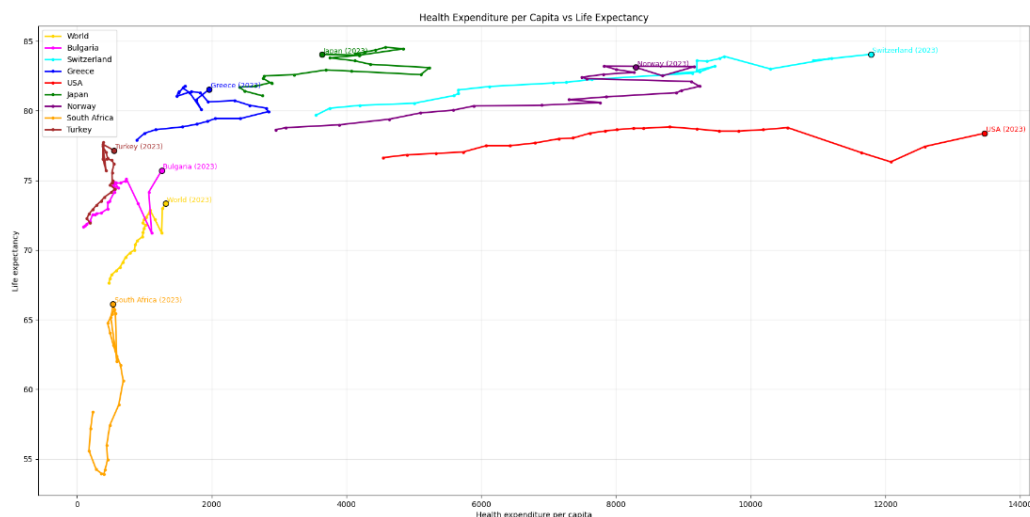


Fig. 1. Health Expenditure per Capita vs Life Expectancy.

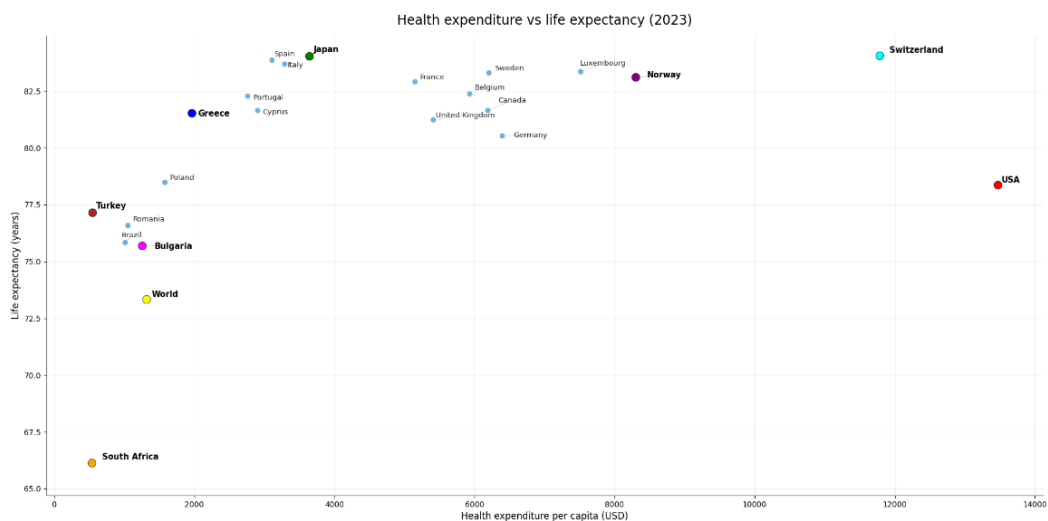


Fig. 2. Health Expenditure vs Life Expectancy (2023).

These graphs reflect that while there is generally a strong correlation between the two variables, it is not absolute, and furthermore, it is noted that there is not necessarily a causal relationship between them. Some of the reasons why life expectancy is not directly linked to costs are that it also depends on various parameters, which will be examined below.

2.2. Correlation analysis

Since all missing values in the specified columns have been handled, a correlation heatmap was generated for the numerical variables in the DataFrame. This will help visualize the relationships between the various features. This visualization presents the Pearson correlation coefficients between all numerical features, as shown in Fig. 3 below.

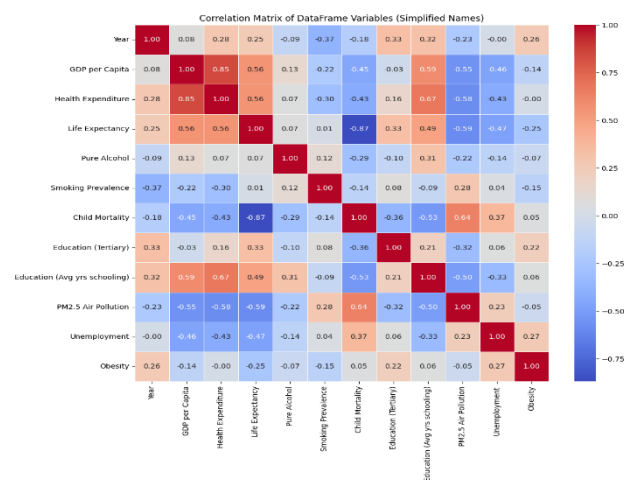


Fig. 3. The Correlation Matrix.

Based on the above correlation matrix, the strongest positive and negative correlations can be observed:

The strongest positive correlation observed is between 'GDP per capita' and 'Health expenditure', with a correlation coefficient of 0.85. This indicates a very strong direct relationship, suggesting that countries with higher GDP per capita tend to also have higher health expenditure.

On the other hand, the strongest negative correlation is between 'GDP per capita' and 'Hospital beds per 1000 people' with a correlation coefficient of -0.20. This suggests a weak inverse relationship, where countries with higher GDP per capita tend to have slightly fewer hospital beds per 1000 people. However, it is important to note that this is a relatively weak correlation. To provide a clearer overview, a bar chart was created showing the five strongest correlations, clearly illustrating the most significant positive and negative relationships within the dataset.

In the chart in Fig. 4, each bar represents a pair of variables. It is easy to be noticed the high negative correlation between life expectancy and child mortality, and conversely, the high positive correlation between GDP, health expenditure and education. This suggests that economic prosperity and educational attainment are often linked to higher health spending.

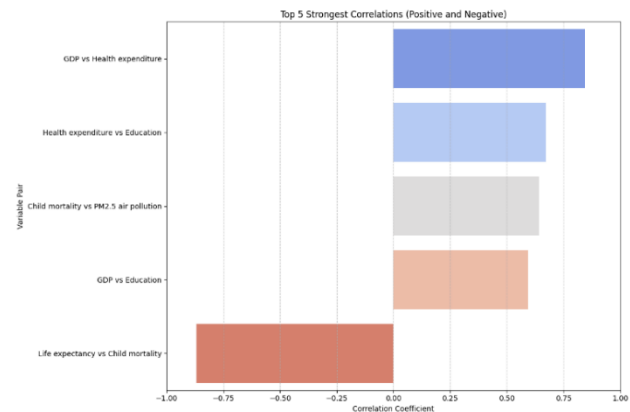


Fig. 4. The Correlation Bar Chart.

Subsequently, the correlation coefficients between 'Life Expectancy' and all other variables in the dataset were identified (Table 1).

Table 1. Correlations with 'Life Expectancy'.

Variable	Correlation
Life Expectancy	1.000000
Health Expenditure	0.561414
GDP per Capita	0.558794
Education (Avg yrs schooling)	0.493118
Education (Tertiary)	0.333733
Year	0.249680
Pure Alcohol	0.070023
Smoking Prevalence	0.011482
Obesity	-0.246621
Unemployment	-0.474428
Air Pollution	-0.589279
Child Mortality	-0.868321

The main conclusions that can be drawn from the Table 1 above are as follows:

- Firstly, the strongest positive correlations of Health Expenditure (0.561) and GDP per Capita (0.559), indicates that higher health expenditure is moderately associated with higher life expectancy and respectively, countries with higher GDP per capita, tend to have higher life expectancy. In addition, the education, more years of schooling, is linked to increased life expectancy (0.493).

- The strongest negative correlations are between child mortality (-0.868), indicating that a higher child mortality rate is strongly associated with lower life expectancy. This is a very significant relationship. Furthermore, PM2.5 air pollution (-0.589) indicates that higher levels of air pollution are moderately associated with lower life expectancy. The same principle applies to unemployment, (-0.474), and higher unemployment rates are linked to lower life expectancy.

- In summary, factors like child mortality, air pollution, and unemployment have a significant negative impact on life expectancy, while health expenditure, GDP per capita, and education show a strong positive relationship.

- Finally, weak or negligible correlations can be observed in pure alcohol (0.070) and smoking prevalence (0.011). These weak positive correlations, may be due to confounding factors or the specific nature of the dataset.

The ten strongest correlations with life expectancy, are presented in the following Fig. 5.

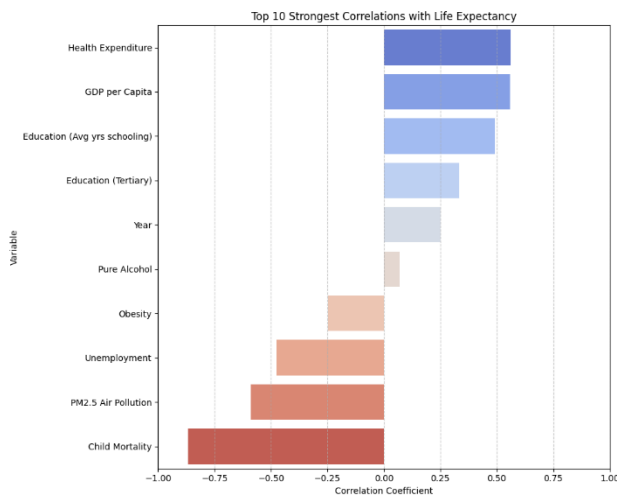


Fig. 5. The 10 strongest correlations with life expectancy.

The above heatmap provides a comprehensive overview of the linear relationships within the specific dataset. As a continuation of this effort, a regression analysis will be conducted, to examine how various variables quantitatively predict life expectancy.

2.3. Regression analysis

As mentioned above, regression analysis was implemented, to understand how different variables quantitatively predict 'Life Expectancy'. For this reason, Multiple Linear Regression, was performed. The model was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared, and the coefficients were analyzed. Upon completion of the process, the main results are as follows:

- **Model Evaluation:** The MSE was found 0.02. This low MSE indicates that the model's predictions are close to the actual values. Moreover, the RMSE, which represents the typical error magnitude, was found 0.16. Finally, the R-squared (R^2) was found 0.82. This means that approximately 82% of the variance in 'Life Expectancy' can be explained by the features included in the model, which is a relatively strong R-squared, suggesting a good fit.

- **Model Coefficients:** The Intercept, the expected 'Life Expectancy' when all other predictor variables are zero, was found 0.05. Furthermore, the Child Mortality of -0.82, is the strongest negative predictor. For every one-unit increase in 'Child Mortality', 'Life Expectancy' is predicted to decrease by 0.82 units, holding other variables constant. This aligns with our earlier correlation findings. The Unemployment of -0.53, suggests that one-unit increase in 'Unemployment' is associated with a 0.53-unit decrease in 'Life Expectancy'.

The variables Pure Alcohol (-0.28), Obesity (-0.17), PM2.5 Air Pollution (-0.12), Smoking Prevalence (-0.12), and GDP per Capita (-0.18), also show negative coefficients, meaning an increase in these factors is associated with a decrease in life expectancy. The negative coefficient for GDP per Capita might seem counter-intuitive at first, given its strong positive correlation. This

often happens in multiple regression due to multicollinearity or when the effect of a variable is conditional on other variables in the model. It suggests that once the effects of other variables like Health Expenditure and Education are accounted for, GDP per Capita might have a different marginal effect.

The Education (Avg years schooling) also has a negative coefficient (-0.06), which is again counter-intuitive. As a result these relationships should need further investigation.

Finally, the Health Expenditure of 0.05 is the only variable with a positive coefficient. A one-unit increase in 'Health Expenditure' is predicted to increase 'Life Expectancy' by 0.05 units. The overall results are presented in the following Table 2.

Table 2. Regression Analysis-Model Coefficients.

Feature	Coefficient
0 GDP per Capita	-0.184001
1 Health Expenditure	0.045824
2 Child Mortality	-0.823087
3 Education (Avg yrs schooling)	-0.061094
4 Air Pollution	-0.120481
5 Unemployment	-0.531425
7 Pure Alcohol	-0.275931
6 Obesity	-0.171990
8 Smoking Prevalence	-0.118452

Moreover, the Variance Inflation Factor (VIF) values have been calculated to assess multicollinearity among the predictor variables:

- Health Expenditure (122.96), GDP per Capita (51.81), PM2.5 Air Pollution (36.32), Child Mortality (22.11), and Education (Avg yrs schooling) (19.41) all show very high VIF values (significantly above the common threshold of 5 or 10). This indicates a strong presence of multicollinearity among these variables.

- Unemployment (8.64) is also above the common threshold of 5, suggesting a moderate level of multicollinearity.

- Smoking Prevalence (4.24), Pure Alcohol (3.40), and Obesity (2.18) have VIF values below 5, indicating they are less affected by multicollinearity in this model.

The impact on negative coefficients can be a high multicollinearity which can make the interpretation of individual regression coefficients unstable and potentially misleading. When predictor variables are highly correlated, the model struggles to isolate the unique effect of each variable on the dependent variable. This can lead to unexpected positive or negative signs, inflated standard errors, and the overall statistical power of the model might decrease [5,6]. In the current model, the high VIF values for GDP per Capita, Health Expenditure, Child Mortality, Education (Avg yrs schooling), and PM2.5 Air Pollution suggest that their coefficients (some of which are negative) should be interpreted with caution. So, it's challenging to precisely determine the independent effect of each of these highly correlated variables on 'Life Expectancy'. To address this, Principal Component Analysis (PCA) will be used.

2.4. Principal component analysis

Based on the high VIF scores of the previous chapter, PCA was applied to the following highly correlated features: 'Health Expenditure', 'GDP per Capita', 'PM2.5 Air Pollution', 'Child Mortality', and 'Education (Avg yrs schooling)'. After these features were scaled, the Principal Component Analysis has been successfully applied to the selected highly correlated features. As a result they were transformed into a smaller set of principal components. The output gave the following variance ratio:

- The PC1 (Principal Component 1) explains 65.63% of the variance in the original five features.
- The PC2 (Principal Component 2) explains an additional 15.35% of the variance.

Together, PC1 and PC2 explain approximately 81% of the total variance in the original PCA features. This means that by using just these two components, we retain a significant amount of information from the original five features. The loadings for PC1 and PC2, are presented to the following Table 3.

Table 3. The loadings for PC_1 and PC_2.

Feature	PC1	PC2
Health Expenditure	0.489190	0.403190
GDP per Capita	0.474301	0.380021
Air Pollution	-0.429355	0.439111
Child Mortality	-0.394383	0.683411
Education (Avg yrs schooling)	0.442547	0.182077

The interpretations from the above are as follows:

- For PC1, it appears to be a general 'Socio-Economic Development and Health Access' factor. Higher values of PC1 would correspond to higher GDP, more health expenditure, better education, and lower air pollution and child mortality. This component captures the general well-being and development aspects of a country.
- For PC2, it seems to be largely driven by 'Child Health and Environmental Quality'. Higher values of PC2 would primarily indicate higher child mortality and higher air pollution, even if other factors like health expenditure and GDP also contribute positively but less strongly. This component might distinguish countries facing significant challenges in child health and environmental quality, possibly independent of the overall development captured by PC1.

The variance ratio of the rest are PC3: 0.1001, PC4: 0.0633, and PC5: 0.0268. The number of chosen_components was 2, meaning the five original highly correlated features have been replaced by two new principal components: PC1 and PC2.

The X train transformed and X test transformed dataframes now have a reduced number of columns, and include PC1 and PC2 alongside the features that were not part of the PCA (e.g., 'Unemployment', 'Obesity', 'Pure Alcohol', 'Smoking Prevalence'). This transformation has effectively addressed the multicollinearity issue by creating uncorrelated principal components.

The next step would be to re-run the Linear Regression model using these transformed datasets to see how the model performance and coefficients change.

2.5. Linear Regression model (using the PC1 and PC2 variables)

The Since PCA has created a cleaner environment for the other variables to express their individual relationships with 'Life Expectancy', and this provided a more stable and interpretable regression model. The results of this new model are:

For Model Evaluation: MSE 7.02 and for RMSE 2.65. On average, the model's predictions for 'Life Expectancy' are off by about 2.65 years, which is a relatively small error given the scale of life expectancy values. Finally the R-squared is 0.82, and this indicates that approximately 82% of the variance in 'Life Expectancy' can be explained by the features included in the model (the two principal components and the remaining original features). This is a strong fit, suggesting the model is good at predicting life expectancy.

The Model Coefficients are shown in Table 4.

Table 4. New Regression Analysis-Model Coefficients.

Feature	Coefficient	Feature	Coefficient
PC1	2.04	Unemployment	-0.10
PC2	-2.27	Pure Alcohol	-0.42
Smoking Prevalence	0.39	Obesity	-0.16

It follows the interpretation of coefficients:

- For PC1 (Coefficient: 2.04): A one-unit increase in PC1 is associated with an increase of approximately 2.04 years in 'Life Expectancy', holding other factors constant. This is a strong positive relationship.
- PC2 (Coefficient: -2.27): A one-unit increase in PC2, representing 'Child Health and Environmental Quality', is associated with a decrease of approximately 2.27 years in 'Life Expectancy', holding other factors constant. This is a strong negative relationship. A higher PC2 score actually represents worse child health and higher air pollution, which would logically lead to lower life expectancy, thus explaining the negative coefficient.
- Smoking Prevalence (Coefficient: 0.39): This still shows a positive coefficient, which remains counter-intuitive. It suggests that a one-unit increase in smoking prevalence is associated with a 0.39-year increase in life expectancy, which contradicts general health knowledge. This could indicate a complex interaction not fully captured by the model or other confounding factors.
- Unemployment (Coefficient: -0.10): A one-unit increase in 'Unemployment' is associated with a decrease of 0.10 years in 'Life Expectancy', which aligns with expectations.
- Obesity (Coefficient: -0.16): A one-unit increase in 'Obesity' is associated with a decrease of 0.16 years in 'Life Expectancy', which is also expected.
- Pure Alcohol (Coefficient: -0.42): A one-unit increase in 'Pure Alcohol' consumption is associated with a decrease of 0.42 years in 'Life Expectancy', which is consistent with known health impacts.
- Intercept (86.44): This is the predicted 'Life Expectancy' when all features are zero. In this context, with scaled and transformed features, it doesn't have a direct interpretable meaning on its own but is part of the regression equation.

Finally, a scatter plot was generated to visualize the model's predictions against the actual values of Life Expectancy.

In the Fig. 6, the red dashed line represents a perfect prediction scenario, where predicted values exactly match actual values. It can be observed that most of the points cluster closely around the red dashed line, especially in the middle range of life expectancy. This indicates that the model generally makes accurate predictions for a significant portion of the data.

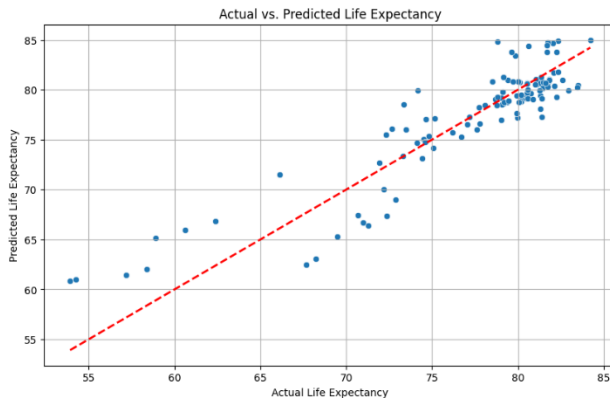


Fig. 6. Actual vs. Predicted Life Expectancy.

On the other hand, there's some spread around the line, particularly at the lower and higher ends, suggesting that the model might have slightly less accuracy for extreme life expectancy values. However, considering the R-squared of 0.82, the model provides a strong fit overall.

In order to improve forecasts, especially for extreme values, a Random Forest regressor model we trained.

2.6. Random Forest model

Random Forest is a powerful model capable of capturing nonlinear relationships and interactions in the data [7]. The model was successfully trained, and its evaluation shows a significant improvement over the linear regression model.

The Random Forest achieved an R-squared of 0.98, which is much higher than the 0.82 of the linear regression. This means that 98% of the variability in 'Life Expectancy' is explained by the Random Forest model.

Furthermore, the RMSE decreased significantly from 2.65 to 0.87. This indicates that, on average, the Random Forest model's predictions are less than 1 year off from the actual values, which is an excellent improvement.

These results confirm that the Random Forest model is much more accurate in predicting 'Life Expectancy,' especially for outliers, due to its ability to capture non-linear relationships and interactions.

Also, the feature importances are PC1=0.65, PC2=0.02, Unemployment=0.20, Obesity=0.07, Pure Alcohol =0.04, and Smoking Prevalence=0.02.

The PC1, representing socioeconomic development and access to healthcare, is by far the most important feature for predicting 'Life Expectancy' in the Random Forest model, as well as in linear regression.

Unemployment and obesity follow in importance, while PC2, alcohol consumption, and smoking have a

smaller effect in the Random Forest model. Fig. 7 scatter plot visualize the model's predictions.

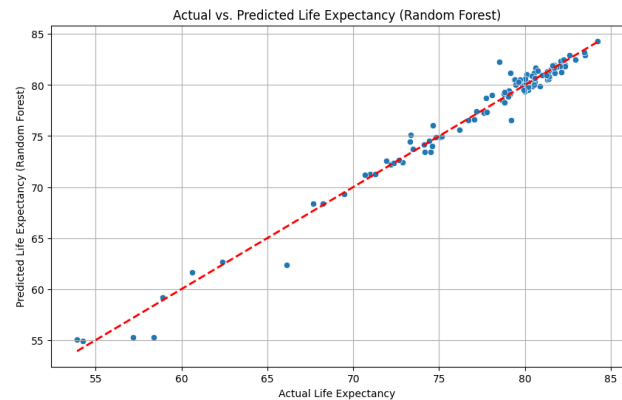


Fig. 7. Actual vs. Predicted Life Expectancy(Random Forest).

The final 'Actual vs. Predicted Life Expectancy (Random Forest)' chart shows that the data points are now much closer to the red dashed line across the entire range of values, including the outliers.

3. Conclusion

This study has completed an extensive analysis of the data, from data cleaning to the training and evaluation of two different regression models for predicting life expectancy.

Initially, the investigation of correlations revealed strong relationships between variables, with the strongest positive correlation to be observed between 'GDP per capita' and 'Health expenditure' (0.85), indicating that countries with higher GDP tend to have higher health expenditures.

With regard to correlations with Life Expectancy, strong negative was observed with 'Child Mortality' (-0.87) and 'PM2.5 Air Pollution' (-0.59). This is to be expected, as high child mortality and poor air quality are linked to reduced life expectancy. In contrast, Life Expectancy showed a strong positive correlation with 'Health Expenditure' (0.56) and 'GDP per Capita' (0.56), highlighting the importance of economic prosperity and health investments.

The revised linear model with PCA had an R-squared of 0.82 and RMSE of 2.65, suggesting a good level of prediction, though with some weakness in extreme values.

Random Forest Regressor was implemented to further improve predictions, using the same variables, including PC1 and PC2. The Random Forest model demonstrated excellent improvement, with R-squared 0.98, and RMSE 0.87.

The overall analysis highlights the strong impact of socio-economic development, health investments, child mortality, and environmental quality on life expectancy. While linear regression provided a good initial understanding, the complexity of the relationships required a more robust model. The Random Forest Regressor proved to be highly effective, providing a very accurate and reliable model for predicting life expectancy, successfully addressing both general trends and extreme values in our data.

Some approaches for further investigation of the issue could involve the use of polynomial regression, residual analysis, weighted regression, and the enrichment of the data, in terms of the number of variables and their quality.

References

- [1] Dattani, S. Rodés-Guirao, L. Ritchie, H. Ortiz-Ospina, E. & Roser, M. (2023) Life Expectancy. Published online at OurWorldinData.org., <https://ourworldindata.org/life-expectancy>.
- [2] Gazilas, E. T. (2024) Factors Influencing Life Expectancy in Low- Income Countries: A Panel Data Analysis. *Journal of Applied Economic Research*.
- [3] World Bank's open data platform. Available online:<https://www.worldbank.org/ext/en/home> (accessed on 10 April 2026).
- [4] Malarvizhi, R. Thanamani, A. (2012) K-Nearest Neighbor in Missing Data Imputation. *International Journal of Engineering Research and Development* . Volume 5, Issue 1.
- [5] Tsagris, M. Pandis, N. (2021) Multicollinearity. *Statistics and research design*. Volume 159, Issue 5, 2021.
- [6] Shrestha, N. (2020) Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics*, Vol. 8, No. 2, 39-42.
- [7] Auret, L. Aldrich, C. (2012) Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, Volume 35, Pages 27-42.

Security Considerations and Techniques for Video Streaming Distribution in Home ISPs

Iliyan Iliev ^{1,*}, Ivan Blagoev ¹ and Krassimira Stoyanova ¹

¹ Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences, Acad. G. Bonchev St., 1113 Sofia, Bulgaria; ivan.blagoev@iict.bas.bg, krassimira.stoyanova@iict.bas.bg

*E-mail of corresponding author: iliyen.iliev@iict.bas.bg

Abstract: The increasing adoption of Internet live-streaming services and high-resolution video formats generates significant network traffic and requires stable high-bandwidth Internet connectivity. However, many regional Internet service providers (ISPs) still face practical limitations in delivering high-quality live-streaming services to subscribers, particularly in environments with limited external Internet capacity. This paper proposes an architectural approach for secure redistribution of Internet live-streaming services within local ISP infrastructures based on passive optical network (PON) environments. The proposed model combines multicast-oriented traffic distribution with encrypted HTTP Live Streaming (HLS), conditional access mechanisms, certificate-based authentication, and dynamic cryptographic key management for IPTV set-top boxes. In addition, the paper discusses security considerations related to authentication, encrypted key delivery, band-width optimization, and protection against unauthorized redistribution of multimedia content.

Keywords: cybersecurity; live streaming; IPTV; FTTH; PON; multicast; HLS; video encryption; conditional access; authentication; content protection.

Received 30 March 2026

Accepted 5 July 2026

Published 23 September 2026

1. Introduction

The rapid growth of Internet-based live-streaming services and over-the-top (OTT) platforms has significantly increased the demand for reliable high-bandwidth multimedia delivery [1]. Modern high-resolution video formats, including 4K and High Frame Rate (HFR) streaming, require stable network connectivity and efficient traffic distribution mechanisms despite improvements in video compression technologies [2]. In regional and remote areas, Internet service providers (ISPs) may still experience limitations related to external Internet capacity and scalability of multimedia delivery infrastructures.

Passive optical network (PON) technologies are increasingly deployed in ISP environments due to their efficient bandwidth utilization and scalability [3]. Such infrastructures create opportunities for implementing local redistribution mechanisms for live-streaming services in order to reduce external traffic consumption and improve quality of service for end users. In this context, multicast-based delivery models are particularly suitable because they minimize unnecessary traffic replication during simultaneous distribution of identical multimedia streams to multiple subscribers [4,5].

At the same time, redistribution of protected multimedia content introduces important cybersecurity

and content-protection challenges. Modern streaming platforms rely on encryption, authentication, and conditional-access mechanisms to prevent unauthorized access and illegal redistribution of copyrighted content. HTTP Live Streaming (HLS), one of the most widely adopted streaming technologies, supports encrypted MPEG Transport Stream delivery using AES-based protection mechanisms defined in RFC 8216 [7].

This paper proposes an architectural approach for secure redistribution of Internet live-streaming services within local ISP environments. The proposed model combines multicast-based distribution in PON infrastructures with encrypted HLS delivery, external authentication services, and dynamic cryptographic key management for IPTV set-top boxes (STBs). The paper additionally discusses practical considerations related to conditional access, authentication workflows, encrypted key delivery, and protection against unauthorized redistribution of multimedia content.

1.1. Main Contributions

This paper proposes an architectural approach for secure redistribution of Internet live-streaming services within ISP-operated passive optical network environments. The proposed model combines multicast-oriented traffic distribution with encrypted HTTP Live Streaming (HLS), certificate-based authentication, and dynamic cryptographic key management for IPTV set-top boxes [6–8].

Particular attention is given to the separation between content distribution, authentication services, and key-management infrastructures in order to reduce direct



exposure of protected multimedia content and sensitive user information. The proposed workflow additionally addresses practical considerations related to bandwidth optimization, secure key delivery, and prevention of unauthorized redistribution in local ISP environments [3–5].

The presented architecture is intended as a practical engineering framework that may support future implementation and performance evaluation of secure live-stream redistribution systems in real ISP infrastructures.

2. Proposed Architecture for ISP Live Streaming Redistribution

The proposed approach is based on the deployment of a local streaming node within the infrastructure of an Internet service provider (ISP) for the purpose of redistributing Internet live-streaming content to end users independently of limitations in their external Internet connectivity. Such an architecture aims to improve the accessibility of high-quality protected multimedia services in regional ISP environments while simultaneously reducing the amount of external traffic generated by multiple subscribers consuming identical live-streaming content.

The implementation of local redistribution mechanisms may provide several practical advantages for Internet service providers and their subscribers. By maintaining local synchronization of multimedia streams and redistributing them through the ISP infrastructure, subscribers may access high-quality protected video content with improved stability and reduced dependence on external network conditions. At the same time, the reduction of replicated external traffic may improve bandwidth efficiency for the provider, particularly during large-scale live broadcasts involving a significant number of simultaneous viewers. In addition, the deployment of local redistribution platforms may create opportunities for promoting regional multimedia initiatives and locally distributed video services.

The proposed architecture is primarily focused on deployment within passive optical network (PON) environments, as PON technologies are increasingly replacing traditional copper-based access infrastructures such as ADSL, VDSL, and copper Ethernet networks. Modern PON deployments provide scalable bandwidth distribution models and support efficient integration of multimedia services within ISP infrastructures [3]. In such environments, multicast-oriented traffic distribution mechanisms become particularly suitable because they minimize unnecessary replication of identical multimedia traffic and improve scalability during simultaneous delivery to multiple subscribers.

The implementation of secure local redistribution of protected multimedia content requires consideration of several key technical aspects. These include efficient traffic distribution techniques capable of avoiding unnecessary concentration of traffic at a single network point, secure authentication and encryption mechanisms for protecting multimedia streams, and additional measures intended to reduce the risk of unauthorized redistribution of copyrighted content.

Modern web-streaming services typically rely on separate protocols for upstream media contribution and downstream content delivery, with RTMP and HTTP Live Streaming (HLS) being among the most widely adopted protocol combinations. This architectural separation has evolved as a practical response to the characteristics of consumer Internet connectivity and the requirements of adaptive multimedia delivery. Within the HLS protocol, multimedia streams are divided into encrypted video segments accompanied by periodically updated M3U playlists describing the available media fragments and corresponding decryption parameters [7]. To achieve service reliability, low latency, and efficient load balancing, modern streaming platforms are commonly deployed using content delivery network (CDN) architectures consisting of geographically distributed streaming nodes.

In the proposed model, the local ISP streaming node operates as an intermediary redistribution component between the external streaming platform and the local subscriber network. The node continuously synchronizes encrypted multimedia segments from the original streaming service while simultaneously redistributing the content to local subscribers through the ISP infrastructure. This approach allows optimization of external bandwidth consumption while preserving compatibility with existing streaming technologies, authentication workflows, and content protection mechanisms.

3. Distribution techniques

3.1. Unicast Combined with Server–Client Communication

Unicast server–client communication remains the most widely used approach for Internet video streaming due to its simplicity, reliability, and compatibility with existing web infrastructures. Modern streaming platforms commonly rely on unicast delivery because it allows straightforward session management and efficient integration with adaptive streaming technologies such as HTTP Live Streaming (HLS) and content delivery network (CDN) architectures [7]. However, in local ISP redistribution environments, pure unicast delivery leads to repeated transmission of identical multimedia traffic between the streaming source and the local redistribution node. As a result, the approach is less suitable in scenarios where minimization of replicated external traffic and bandwidth optimization are primary objectives.

3.2. Peer-to-Peer (P2P)

Peer-to-Peer (P2P) communication represents a distributed traffic-delivery approach intended to reduce concentration of network traffic at a single distribution point. P2P mechanisms have historically been applied in file-sharing systems, multimedia distribution platforms, device discovery protocols, and decentralized communication services. In multimedia streaming environments, P2P architectures may improve scalability by distributing traffic generation among participating client devices instead of relying exclusively on centralized streaming servers.

Despite these advantages, large-scale P2P redistribution is less suitable for passive optical network (PON) infrastructures operating in Point-to-Multipoint (P2M) environments. In such networks, multiple optical network terminals (ONTs) share common downstream and upstream bandwidth resources associated with a single optical line terminal (OLT) port. Consequently, uncontrolled peer-generated traffic may introduce additional network congestion and reduce the efficiency of bandwidth utilization, particularly in scenarios involving high-bitrate live-streaming services.

3.3. Multicast

Multicast communication provides an efficient mechanism for distributing identical multimedia traffic simultaneously to multiple subscribers without unnecessary replication at a single network point. For this reason, multicast technologies are particularly suitable for live-streaming services deployed within modern passive optical network infrastructures [4,5]. In multicast environments, a single multimedia stream may be transmitted through the network and replicated only at the appropriate distribution points, thereby significantly reducing bandwidth consumption compared to traditional unicast delivery models.

The application of multicast delivery in PON environments extends traditional multicast concepts used in enterprise and carrier-grade networks by introducing additional mechanisms associated with multiplexed optical transmission and GEM frame processing [5,6]. In such architectures, multicast frames may contain identifiers corresponding to the destination optical network terminals, allowing efficient distribution of multimedia traffic between the optical line terminal and multiple subscriber devices. The general multicast distribution model in a GPON environment is illustrated in Figure 1:

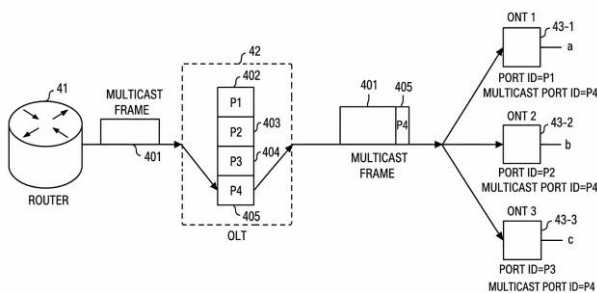


Figure 1. Multicast GPON transmission and GEM frame processing method multicast frame transmitted in PON (adapted from [6]).

4. Security Considerations

4.1. Communication security

Communication security in multimedia streaming systems is traditionally implemented through conditional-access architectures similar to those used in Digital Video Broadcasting (DVB) ecosystems. In such systems, cryptographic key exchange procedures and content scrambling mechanisms are logically separated, while specific conditional-access implementations, including

Conax, Irdeto, VideoGuard, and Viaccess, integrate these functions into unified content-protection frameworks [8,9]. The primary objective of these mechanisms is to restrict access to protected multimedia content exclusively to authorized client devices and authenticated subscribers.

A comparable security model may be applied to HTTP Live Streaming (HLS)-based infrastructures. The HLS specification defined in RFC 8216 supports encryption of MPEG Transport Stream (TS) video segments using AES-based cryptographic protection mechanisms [7]. In this approach, multimedia content is encrypted on the streaming server before distribution, while authorized client devices obtain the information necessary for decryption through standardized M3U playlists and associated key-delivery procedures. The playlist metadata specifies the encryption method, cryptographic parameters, and the location of the corresponding decryption keys required for processing protected video segments.

Such an architecture allows encrypted multimedia traffic to be distributed through intermediate network infrastructures, including ISP-operated redistribution nodes, without exposing the original video content in unencrypted form. As a result, the streaming infrastructure may maintain compatibility with existing CDN and multicast-oriented delivery models while preserving secure conditional access to protected multimedia services.

4.2. Authentication

In encrypted multimedia streaming environments, the protection of video segments alone is insufficient without reliable mechanisms for authenticating authorized client devices and subscribers. Since encrypted HLS segments cannot be decrypted without access to the corresponding cryptographic keys, authentication procedures may be focused primarily on controlling secure access to key-delivery services rather than to the multimedia traffic itself [8,10]. Such an approach allows encrypted video content to be distributed through intermediate infrastructures while restricting decryption capabilities exclusively to authorized playback devices.

Previous research has demonstrated that conditional-access mechanisms for web-based video streaming may be implemented through authenticated user sessions and cryptographic token exchange between client devices and streaming services [10]. A similar approach may be applied within the proposed ISP redistribution architecture, where authentication procedures are integrated directly into the cryptographic key-delivery workflow instead of the multimedia segment delivery process itself. In this model, the streaming infrastructure distributes encrypted multimedia traffic independently of the authentication subsystem, while access to decryption keys remains protected through authenticated sessions and authorization tokens.

The proposed authentication architecture relies on interaction between IPTV set-top boxes (STBs), external trust-service providers, and key-management infrastructures responsible for distributing authorized cryptographic material. After successful user authentication, the client device obtains temporary

authorization credentials that allow secure retrieval of the corresponding decryption keys required for processing protected multimedia segments. Such separation between encrypted multimedia delivery and authentication procedures improves scalability of the streaming infrastructure while reducing direct exposure of protected content and sensitive authentication information.

In practical ISP environments, encrypted HLS playback is expected to be performed primarily through dedicated IPTV set-top boxes rather than general-purpose web browsers. Consequently, the authentication workflow must support automated session establishment, secure token management, and controlled cryptographic key distribution within consumer multimedia devices. The overall architecture of the proposed redistribution model, including streaming infrastructure, authentication services, and cryptographic key management, is illustrated in (Figure 2).

4.3. Example Workflow for System Operation

The proposed redistribution architecture relies on integration between IPTV set-top boxes (STBs), external trust-service providers, and cryptographic key-management infrastructures responsible for authorizing access to protected live-streaming services. Initially, the subscriber registers with a trust-service provider integrated into the streaming ecosystem and creates a user profile accessible through a mobile application or similar authentication platform. In practical deployments, such identification procedures are typically performed remotely through secure online verification mechanisms.

When a protected live-streaming event becomes available, the subscriber selects the corresponding streaming service and associated trust-service provider directly from the IPTV set-top box interface. The STB subsequently generates an asymmetric cryptographic key pair, where Elliptic Curve Cryptography (ECC) is

recommended because it provides comparable security strength with substantially shorter key sizes than RSA [14]. The device additionally generates a PKCS#10 Certificate Signing Request (CSR) containing a randomly generated identifier, such as a UUID-based common name, intended to minimize direct exposure of user-identifiable information.

To establish the authentication workflow, the STB encodes a dedicated authorization URL as a QR code displayed on the television screen. In parallel, the device transmits the generated CSR to the trust-service provider and waits for certificate issuance. The subscriber scans the QR code using the authentication provider's mobile application and confirms authorization for access to the selected live-streaming service. Following successful authentication and authorization, the trust-service provider signs the CSR, issues a temporary X.509 certificate with limited validity corresponding to the duration of the live event, and forwards the certificate information to the streaming infrastructure.

After certificate issuance, the IPTV set-top box downloads the corresponding X.509 certificate and initiates the protected streaming workflow. The streaming platform encrypts the AES decryption keys associated with the multimedia stream using the public keys of authorized subscriber certificates and stores the encrypted key material within a dedicated key repository. To retrieve the required decryption keys, the STB establishes an HTTPS connection to the key-management service using mutual TLS authentication with the issued client certificate and associated private key. Following successful authentication, the key-management infrastructure issues a temporary cryptographic authorization token allowing the device to retrieve the encrypted AES keys required for multimedia playback.

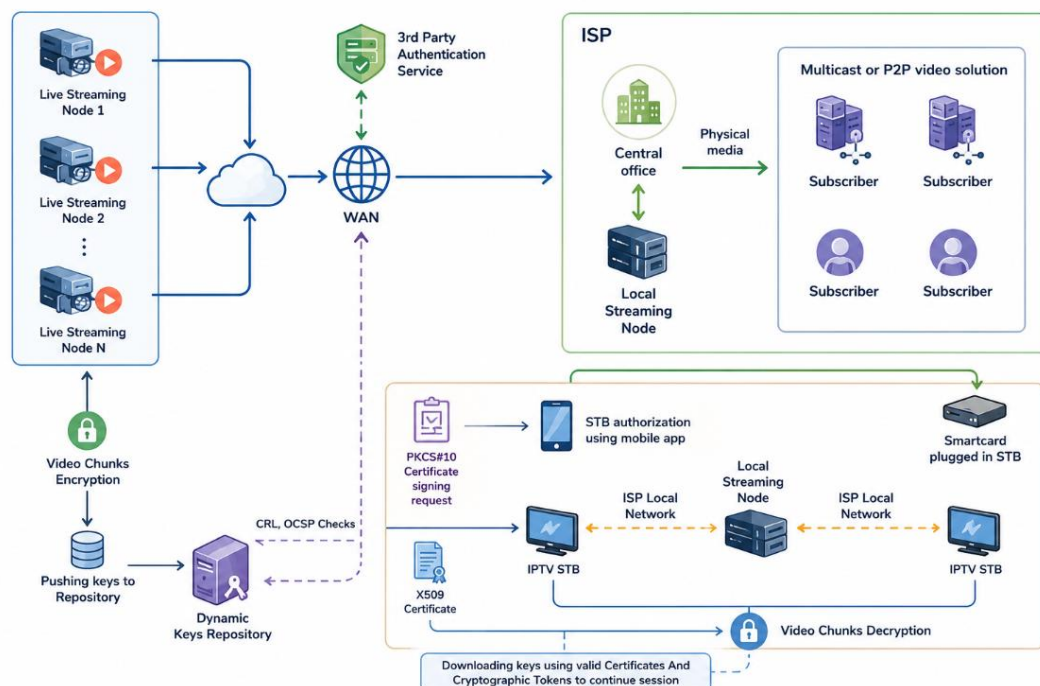


Figure 2. General scheme for Live Streaming redistribution in a local ISP environment.

The IPTV set-top box decrypts the retrieved AES keys locally using its private key and subsequently applies the decrypted symmetric keys to process the encrypted HLS video segments. Once the authenticated session has been established, repeated client-certificate authentication is not continuously required during active playback. Instead, temporary authorization tokens may be periodically renewed within the active session lifecycle, while prolonged inactivity or token expiration requires re-establishment of the authenticated TLS session.

The proposed workflow is designed to minimize unnecessary exposure of personal information and sensitive cryptographic material within the ISP infrastructure. Personal-data processing operations remain isolated within the systems of the external authentication provider, while the ISP infrastructure primarily operates as a transport and redistribution environment for encrypted multimedia traffic. Similarly, the key-management repository may operate as an independent service separated from both the authentication provider and the multimedia distribution infrastructure, thereby reducing direct exposure of decryption capabilities within intermediary network components.

In the presented scenario, X.509 certificates issued by the trust-service provider are used as authentication credentials for multimedia access control. Since typical multimedia streaming services do not require qualified electronic identification procedures, the generated certificates are not necessarily qualified certificates within the meaning of formal electronic identification frameworks. Consequently, private cryptographic keys may be generated and stored directly on consumer devices rather than exclusively within specialized hardware security modules or smart-card environments. While such an approach simplifies deployment and improves usability, it also introduces potential security considerations associated with unauthorized extraction or replication of cryptographic material from compromised playback devices, which remains relevant in the context of illegal redistribution of protected multimedia content.

4.4. Limitations and Performance Overhead

Despite the advantages of the proposed architecture regarding bandwidth optimization and secure redistribution of live-streaming content, several practical limitations and performance considerations must be taken into account during implementation. The integration of authentication services, encrypted key delivery, and conditional access mechanisms inevitably increases the complexity of the overall streaming infrastructure and introduces additional processing and communication overhead.

One of the primary factors affecting system performance is the authentication and session establishment procedure between IPTV set-top boxes, the trust service provider, and the local streaming infrastructure. The proposed workflow relies on asymmetric cryptographic operations, certificate issuance, HTTPS communication, and token-based session management. These mechanisms introduce additional computational work associated with asymmetric cryptographic operations, certificate processing, secure

communication exchanges, and token management during session establishment. In large-scale deployments involving simultaneous authentication requests from numerous subscribers, the trust service provider and key management infrastructure may become a bottleneck if appropriate load-balancing and scaling mechanisms are not implemented.

Additional session-establishment latency may also occur during the initialization phase of live-stream playback. Before receiving decryption keys, the client device must complete authentication procedures, establish secure communication channels, and obtain valid authorization tokens. Although the amount of control traffic associated with these procedures is relatively small compared to the video stream itself, the additional communication exchanges and cryptographic operations may increase playback startup time, particularly during high-demand live events with a large number of concurrent users.

Another important consideration concerns the secure storage and management of cryptographic material within IPTV set-top boxes. In the proposed architecture, private keys may be generated and stored directly on client devices. Although this approach simplifies deployment, secure management and storage of cryptographic keys become critical implementation requirements, since compromise of keying material may undermine the protection of encrypted content [11]. Consequently, unauthorized extraction of cryptographic material from compromised devices remains a potential risk factor for illegal redistribution of protected content.

The multicast-based distribution model additionally introduces synchronization and network-management challenges in passive optical network environments. Efficient multicast delivery requires proper configuration of Internet Group Management Protocol (IGMP) mechanisms, multicast forwarding policies, and traffic isolation procedures between optical network terminals [5]. Improper multicast configuration may lead to packet duplication, unnecessary bandwidth consumption, or instability during large-scale live-stream distribution. Furthermore, encrypted HLS delivery combined with multicast redistribution may require careful balancing between segment duration, synchronization intervals, and caching strategies in order to minimize playback interruptions and buffering delays.

The proposed architecture also depends on the availability and reliability of external trust-service providers responsible for authentication and certificate issuance. Temporary unavailability of such services may prevent subscribers from establishing new playback sessions even if the streaming infrastructure itself remains operational. For this reason, practical deployments should consider redundancy mechanisms, distributed authentication infrastructures, and fallback procedures for maintaining service continuity during network failures or maintenance periods.

Finally, although encryption and authentication mechanisms significantly improve protection against unauthorized access, they cannot fully eliminate the possibility of illegal redistribution of multimedia content after successful playback authorization. Once decrypted

video reaches an authorized endpoint device, various recording and retransmission techniques remain technically possible despite the use of watermarking, HDCP, and conditional access technologies [11,12]. Therefore, practical protection of copyrighted multimedia content continues to require a combination of technical safeguards, legal enforcement mechanisms, and operational monitoring procedures.

5. Prevention against Illegal Redistribution

Protection against unauthorized redistribution remains a major challenge in modern multimedia streaming systems. Unlike analogue broadcasting, digital technologies allow multimedia content to be copied and redistributed without degradation of the original quality, creating favorable conditions for large-scale digital piracy [11].

Modern content-protection strategies therefore combine legal and technical mechanisms intended to limit unauthorized redistribution of protected multimedia services. Legal approaches typically rely on contractual agreements and copyright regulations restricting unauthorized copying and retransmission of multimedia content [11].

In addition, streaming infrastructures employ technical protection mechanisms such as digital watermarking and hardware-based content protection. Watermarking techniques embed identifying information into multimedia streams in order to trace unauthorized redistribution, although practical watermarking schemes must remain robust against various manipulation and removal attacks [12]. Another content-protection mechanism is High-bandwidth Digital Content Protection (HDCP), which protects encrypted digital content transmitted between source and receiver devices. Earlier HDCP protocol versions have been shown to contain authentication and key-exchange weaknesses, including impersonation-related attacks [13].

Consequently, complete prevention of illegal redistribution remains technically difficult even in infrastructures implementing encryption, authentication, watermarking, and hardware-based protection simultaneously. Practical multimedia security architectures must therefore balance usability, deployment complexity, and content-protection effectiveness while acknowledging that no existing solution can fully eliminate the risk of unauthorized redistribution after legitimate playback access has been granted.

6. Discussion

Despite the deployment of modern encryption, authentication, and conditional-access mechanisms, complete prevention of unauthorized redistribution of multimedia content remains technically challenging once protected content reaches an authorized playback device. Consumer environments inherently allow various recording and retransmission techniques, including external capture and screen-recording methods.

Consequently, practical multimedia protection architectures must balance security, usability, deployment complexity, and economic feasibility. Excessively complex protection mechanisms may negatively affect

scalability, infrastructure costs, and user experience without fully eliminating the possibility of unauthorized redistribution.

For this reason, contemporary streaming ecosystems typically rely on a combination of technical safeguards, contractual restrictions, monitoring procedures, and legal enforcement mechanisms rather than on a single isolated protection technology.

7. Conclusion

This paper presented an architectural approach for secure redistribution of Internet live-streaming services within local ISP environments. The proposed model combines multicast-based distribution in passive optical network infrastructures with encrypted HTTP Live Streaming delivery, external authentication services, and dynamic cryptographic key management for IPTV set-top boxes. The approach aims to reduce external ISP traffic consumption while maintaining secure access to protected multimedia content.

The presented architecture demonstrates how encrypted video delivery, certificate-based authentication, and conditional-access mechanisms may be integrated into a unified streaming ecosystem while preserving separation between content distribution, authentication services, and cryptographic key management. At the same time, the implementation of such systems introduces practical challenges related to authentication latency, session management, scalability, and protection against unauthorized redistribution of multimedia content despite the use of encryption, watermarking, and hardware-based protection technologies.

Future work may include experimental validation of the proposed architecture in real ISP environments, analysis of multicast scalability in large PON deployments, and evaluation of authentication overhead, key-distribution performance, and quality-of-service parameters during live-streaming delivery.

References

- [1] Tefertiller, A. & Sheehan, K. (2019). TV in the Streaming Age: Motivations, Behaviors, and Satisfaction of Post-Network Television. *Journal of Broadcasting & Electronic Media*, vol. 63, pp. 595–616. doi: 10.1080/08838151.2019.1698233.
- [2] Mackin, A., Zhang, F. & Bull, D.R. (2019). A Study of High Frame Rate Video Formats. *IEEE Transactions on Multimedia* 2019, vol. 21, 1499–1512. doi: 10.1109/TMM.2018.2880603.
- [3] Arpaci, M., Ho, E. & Yen, W. (2001). Providing Integrated Services with Broadband PON. *Proceedings of SPIE*, Vol. 4583, Metro and Access Networks, pp. 20–30. doi: 10.1117/12.445133.
- [4] AMX by HARMAN. (2015). Multicast for Enterprise Video Streaming. White Paper, v.01.2015. Official AMX white-paper page: <https://adn.harmanpro.com/>.
- [5] Yang, W. & Zhang, L. (2011). An Application Research on IPv6 Multicasting Testing Method. *Procedia Engineering*, vol.24, 143–151. doi: 10.1016/j.proeng.2011.11.2617.
- [6] Lim, S.-Y.; Song, J.-Y.; Kim, J.-H.; Kwon, S.-W.; Lee, Y.-S. & Lee, J.-H. (2011). Multicast GPON Transmission and GEM Frame Processing Method. European Patent EP1499155A1, 19 January 2005. <https://patents.google.com/patent/EP1499155A1>.
- [7] Pantos, R. & May, W. HTTP Live Streaming. RFC 8216. Available online: <https://www.rfc-editor.org/info/rfc8216>.

-
- [8] ITU-R. Conditional-Access Systems for Digital Broadcasting. Recommendation ITU-R BT.1852-1, January 2017. Available online: <https://www.itu.int/rec/R-REC-BT.1852-1-201701-I/en>.
- [9] Weinmann, R.-P. & Wirt, K. (2005). Analysis of the DVB Common Scrambling Algorithm. In *Communications and Multimedia Security*; Springer: Boston, MA, USA, 2005; pp. 195–207. doi: 10.1007/0-387-24486-7_15.
- [10] Iliev, I.; Blagoev, I. (2022). An Approach to Improve Web Video Streaming Security and Prevent Personal Data Leakage. *Information & Security: An International Journal*, vol. 53, pp. 78–88. doi: 10.11610/isij.5306.
- [11] Qureshi, A. & Megías, D. (2021). Blockchain-Based Multimedia Content Protection: Review and Open Challenges. *Applied Sciences*, 11, 1. <https://doi.org/10.3390/app11010001>.
- [12] Malanowska, A., Mazurczyk, W., Koohpayeh Araghi, T., Megías, D. & Kuribayashi, M. (2024). Digital Watermarking—A Meta-Survey and Techniques for Fake News Detection. *IEEE Access*, 12, 36311–36345. doi:10.1109/ACCESS.2024.3374201.
- [13] Zhao, J.; Gu, D.; Li, Y. & Cheng, W. (2012) On weaknesses of the HDCP authentication and key exchange protocol and its repair. *Mathematical and Computer Modelling*, 55(1–2), 19–25. <https://doi.org/10.1016/j.mcm.2011.01.005>.
- [14] Barker, E. (2020). Recommendation for Key Management: Part 1—General. NIST Special Publication 800-57 Part 1, Revision 5; National Institute of Standards and Technology: Gaithersburg, MD, USA. doi: 10.6028/NIST.SP.800-57pt1r5.
-

