Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 3, June 2025, pp. 2042~ 2054
ISSN: 2302-9285, DOI: 10.11591/eei.v14i3.9213

A systematic literature review on the use of artificial intelligence for cybercrime rate forecasting

Manuel Martin Morales Barrenechea[1,2], Miguel Angel Cano Lengua[1,3]
[1]Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima, Perú
[2]Facultad de Ingeniería de Redes y Comunicaciones, Universidad Peruana de Ciencias Aplicadas, Lima, Perú
[3]Facultad de Ingeniería de Sistemas e Informática, Universidad Tecnológica del Perú, Lima, Perú

Article Info

Article history:

Received Aug 29, 2024
Revised Oct 23, 2024
Accepted Nov 19, 2024

Keywords:

Artificial intelligence
Cybercrime
Forecast
Machine learning
Techniques

ABSTRACT

Cybercrime has a significant impact on the quality of life and economy of individuals, businesses and countries, and the speed of the increase has made it a pressing issue in today's digital age. This systematic review aims to identify the artificial intelligence models recently developed to forecast the rate of cybercrime and to help authorities and police forces define strategies in the fight against cybercrime. The PRISMA methodology was used with 229 articles retrieved from Scopus, IEEE and Web of Science, of which 30 met the eligibility criteria. The results showed that the traditional machine learning methods random forest, support vector machine (SVM) and logistic regression (LR) excel in their use to forecast cybercrimes by achieving more accurate results among the different methods tested. It was concluded that machine learning methods are, so far, effective in forecasting the rate of cybercrime, with accuracy ratios of up to 99.9%. However, the potential for future research lies in creating new forecasting models such as autoregressive integrated moving average long short term memory (ARIMA-LSTM) proposed in this study to improve the performance and accuracy of cybercrime forecasting.

Corresponding Author:

Manuel Martin Morales Barrenechea
Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos
Av. Carlos Germán Amezaga 375, Lima, Perú
Email: martin.moralesb@unmsm.edu.pe

1. INTRODUCTION

Crime is a global problem. Is a socioeconomic problem committed by criminals that negatively affects the quality of life and economic growth of a country [1]. With the advancement of technology and society increasingly dependent on it [2], crime has expanded its reach digitally, which is known as cybercrime [3]. In this era of the cyber world, cybercrime has been spreading considerably and becoming a major threat [2], which with the COVID-19 pandemic generated an increase in the crime rate [4], even though countries make denoted efforts to adapt and guarantee cybersecurity [5], [6].

The impact of cybercrime is critical. This considers damage and destruction of data, stolen money, theft of personal and financial data, damage to reputation among others [7]. Cybercrime disrupts business operations, resulting in downtime and lost productivity [8]. It is estimated that the cost caused by cybercrime reached 8 trillion dollars in 2023 and will increase 15% annually, reaching 10.5 trillion dollars by 2025 [2], [9]. Beyond the financial and operational effects, cybercrime has significant social implications, such as stress, anxiety and fear among people [10].

Due to their relevance, these impacts demand a comprehensive understanding and strategic response to mitigate the risks associated with cybercrime. Artificial intelligence is increasingly being leveraged in the fight against cybercrime by automating data analysis and improving decision-making processes [11]. Artificial intelligence algorithms are proof items in the field of predictive analytics for fraud detection [12] where researchers have shown that they can work effectively for cybercrimes with an accuracy between 70% and 90% [13].

Artificial intelligence poses great challenges. Various statistical and machine learning techniques have been employed to predict cybercrime rates, including support vector machines (SVM) and k-nearest neighbors (KNN), which have shown promising accuracy rates of up to 89% and 92.34% [14], [15]. Advanced models like bi-directional recurrent neural networks with long short-term memory (BRNN-LSTM) have been proposed to enhance prediction accuracy by accommodating the statistical properties of cyberattack time series data [16]. The integration of these predictive models improves the allocation of resources for cyber defense, aiming to reduce the incidence of cybercrime [17]. As cyber threats evolve, continuous improvement in forecasting methodologies remains crucial for effective cybersecurity strategies [5].

Gaps in cybercrime rate forecasting include limited scope. Existing models do not consider a comprehensive approach to all types of cybercrime [17]. Many studies emphasize the use of machine learning techniques but do not adequately address the integration of diverse datasets or the preprocessing challenges that can affect model accuracy [5], [18]. While some models achieve high accuracy rates, such as 91% with SVM, there remains a need for improved methodologies that can adapt to the evolving nature of cyber threats and incorporate real-time data effectively [18]. Lastly, the reliance on historical data without considering emerging trends in cybercrime presents a significant limitation in predictive capabilities [16].

This study is motivated by contributing to the academic debate on the objective of understanding the progress of predicting the rate of cybercrime using artificial intelligence. The findings of this research have an impact on the actions of the authorities involved in the organized fight against cybercrime. In this regard, evidence is provided, and a new forecasting model autoregressive integrated moving average long short term memory (ARIMA-LSTM) is proposed to improve the performance and accuracy of cybercrime forecasting that could be effectively used by authorities and police forces to formulate cybercrime prevention and control strategies and measures.

The structure of this review is as: section 2 contains the development of this research with the PRISMA methodology. Section 3 presents the results of the questions asked and the discussion of the proposals of the authors considered in this research. In addition, a model proposal ARIMA-LSTM for evaluation by the scientific community. Finally, section 4 presents the conclusion of the findings found in this research.

2. METHOD

This study used the PRISMA methodology. It is justified because it carries out a structured analysis of scientific production [19], [20] by identifying, evaluating, and synthesizing existing studies highlighting trends, methodologies, and gaps [21] to answer the questions raised by this systematic review of the literature [22]. To formulate the questions of this research, the problem, interventions, comparators, and outcomes (PICO) method was used, which allowed specifying the questions and eligibility criteria of the scientific articles of this research [23].

The articles included in this systematic review were classified using the statistical method of hierarchical grouping. It was used to group similar objects into clusters based on their characteristics [24]. The hierarchical grouping made it possible to classify the 30 articles included into four groups according to the modality of cybercrime and the artificial intelligence techniques applied in the forecast. This hierarchical grouping as an instrument facilitated the formulation of [24] of the answers to the research questions posed in this study.

2.1. Research questions

As part of the research process, the PICO method was applied to formulate the main question and four specific research questions (problem, interventions, comparators, and outcomes) with the aim of extracting and summarizing the knowledge of the articles included in this systematic review. These questions are shown in the Table 1.

2.2. Research strategy

As part of the strategy of this study, the search string was constructed following the PICO method, with each of the four defined factors (problem, interventions, comparators, and results) with its respective description, the search terms and a set of synonyms used for this research. The search terms as shown in the Table 2.

Table 1. Research questions

Factor	ID	Question
Main	MQ	What artificial intelligence models to forecast the rate of high-performance cybercrime have been proposed in recent years?
Problem	Q1	What cybercrime related issues have been addressed with the forecast of the rate of cybercrime with artificial intelligence?
Intervention	Q2	What artificial intelligence methods have been used to forecast the rate of cybercrime?
Comparison	Q3	What evaluation metrics have been used to measure the performance of methods for forecasting the rate of cybercrime?
Objective	Q4	What improvements have been incorporated into the models to forecast the rate of cybercrime?

Table 2. Search terms

Factor	Description	Search terms	Synonym
Problem	Cybercrime forecast	"cybercrime rate forecast"	"cybercrime rate forecast", "cybercrime rate prediction", "computer crime rate forecast", "computer crime rate prediction", "cyberfraud rate forecast", "cyberfraud rate prediction"
Intervention	Use of artificial intelligence	"artificial intelligence"	"machine learning", "deep learning", "transfer learning", "reinforcement learning", "neural network"
Comparison	Forecast accuracy	"accuracy"	"precision", "error", "accuracy", "score", "performance", "efficacy", "metric", "statistical"
Objective	Model identification	"models"	"model", "pattern", "method", "algorithm", "technique"

The search terms were combined with Boolean operands with which the following string was constructed and with which the search was performed in the three different databases:
((cybercrime AND rate AND forecast) OR (cybercrime AND rate AND prediction) OR ("computer crime" AND rate AND forecast) OR ("computer crime" AND rate AND prediction) OR (cyberfraud AND rate AND forecast) OR (cyberfraud AND rate AND prediction)) AND ("machine learning" OR "deep learning" OR "transfer learning" OR "reinforcement learning" OR "neural network") AND (precision OR error OR accuracy OR score OR performance OR efficacy OR metric OR statistical) AND (model OR pattern OR method OR algorithm OR technique)

2.3. Eligibility criteria

For this systematic literature review, inclusion and exclusion eligibility criteria were considered to delimit the scope of the search in the databases for articles related to the field of cybercrime rate prediction using artificial intelligence. Three inclusion criteria and three exclusion criteria were considered as shown in the Table 3.

Table 3. Eligibility criteria

Criteria	ID	Question
Inclusion	I1	Articles related to cybercrime
	I2	Articles that apply artificial intelligence in dataset models to forecast the rate of cybercrime
	I3	Journals and conference articles
Exclusion	E1	Articles in languages other than English or Spanish
	E2	Articles published before 2018
	E3	Articles without full text availability

2.4. Information sources

The scientific databases Scopus, IEEE Xplore and Web of Science in Figure 1 were used as a source of information due to their reliability in the academic world. The same search string was used in the three sources of information, obtaining a total of 229 research articles in the period from 2018 to 2024. Among the potentially eligible studies are 140 from Scopus, 83 from IEEE Xplore and 6 from Web of Science.

Figure 1. Information sources

2.5. Article selection process

The article selection process, through the PRISMA flowchart, was carried out in three stages. In the identification stage, the search chain was applied in the databases of the information sources, obtaining the total number of articles that met the defined conditions. In the screening stage, inclusion and exclusion criteria were applied in the title and abstract of the identified articles. In the last inclusion stage, the inclusion criteria were applied at the level of the introduction, method, and conclusions sections to determine the articles that were considered in the qualitative synthesis.

In the identification stage, the search string in the information sources Scopus, IEEE Xplore and Web of Science was applied to determine a total of 229 articles as shown in Figure 2. Then, 19 duplicates were eliminated, leaving 210 studies; In the screening stage, the inclusion and exclusion criteria were applied at the level of title and abstract of the articles, where 24 articles were eliminated to leave 186 documents. In the last stage of inclusion, 156 documents that did not meet any inclusion criteria were eliminated, finally leaving a total of 30 articles for inclusion in the qualitative synthesis.

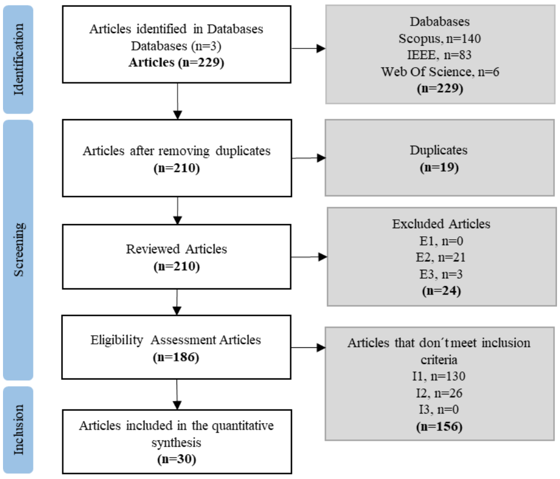

Figure 2. PRISMA flow diagram of the systematic review

2.6. Automatic grouping articles

The grouping of the 30 articles included in this qualitative synthesis was carried out through the hierarchical grouping method. We ensured the objectivity of the analysis and avoided bias in the classification of studies [25]. For the classification, the articles were labeled by characteristics such as the cybercrime modality, the forecasting method used, the type of dataset used, the scaling and imbalance techniques applied to the data, the assembly technique and the optimization of the parameters of the methods used in the models. The DATAtab software was used for hierarchical grouping with Euclidean distance and its graphical representation through a Dendogram as shown in the Figure 3.

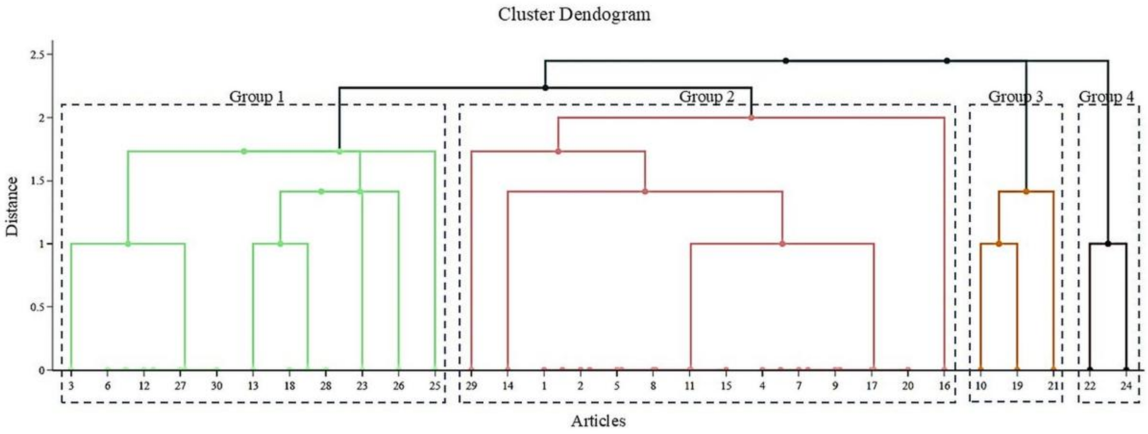

Figure 3. Agglomerative hierarchical grouping of articles

Group 1: composed of eleven documents that address studies on cybercrime, cyberfraud and cyberattack that apply machine learning methods and neural networks. They use time series datasets, scaling techniques to standardize data characteristics, and algorithm assembly techniques [1], [26]-[29], performance optimization with hyperparameters [30], [31] and techniques to handle unbalanced data [32]. Other papers use data without a time series and apply algorithm assembly techniques, techniques for handling unbalanced data, and optimization with hyperparameters [13], [33], [34].

Group 2: composed of fourteen documents that address studies on cyberattack, cybercrime and cyberbullying that apply supervised and unsupervised machine learning methods, and neural networks. They use time series datasets [5] with algorithm assembly techniques [35]-[40]. The rest of the papers use non time series data with algorithm assembly techniques [41]-[45], with data scaling techniques, and optimization with hyperparameters [46], [47].

Group 3: this group is composed of three documents that address studies on cyberattacks, cyberthreats and cyberbullying that use datasets without time series in new hybrid models [48], [49]. Another article uses datasets without time series but applies hyperparameter optimization in forecasting models for these cybercrime modalities [50].

Group 4: this group is composed of two papers addressing cybercrime and cyberfraud studies using time series datasets, scaling techniques to standardize data characteristics, hyperparameter optimization, and algorithm assembly [2]. Another paper considers these optimization techniques but applied to new hybrid models [51].

3. RESULTS AND DISCUSSION

This section answers the research questions posed from the analysis of the 30 articles included in the qualitative synthesis. It is developed in three parts: the first addresses the analysis of the research questions with their respective answers, the second the bibliometric analysis of the selected documents and the third proposes a model based on the findings and opportunities found in the review of the studies.

3.1. Analysis of the research questions

Based on the analysis of the 30 articles included in the qualitative synthesis of this systematic review of the literature, the answers to the questions posed in this research are formulated. For better organization, understanding and support, the answers to the main question and the four specific questions are summarized in tables.

3.1.1. MQ: what artificial intelligence models to forecast the rate of high performance cybercrime have been proposed in recent years?

The artificial intelligence models that have been proposed in recent years to forecast, predict and/or detect different types of cybercrime were classified into four groups using the hierarchical clustering method. This response is detailed in the section on automatic clustering of articles. Next, the review will be further explored through the four specific questions.

3.1.2. Q1: what cybercrime related issues have been addressed with the forecast of the rate of cybercrime with artificial intelligence?

From the articles analyzed, it was found that various problems related to different modalities such as cybercrime, cyberattack, cyberfraud and cyberbullying have been addressed. Artificial intelligence was used through various machine learning methods and neural networks to forecast the cybercrime rate. The response is detailed in the Table 4.

3.1.3. Q2: what artificial intelligence methods have been used to forecast the rate of cybercrime?

From the articles reviewed, it was found that through artificial intelligence, different methods have been used to predict the cybercrime rate. Machine learning methods, bagging methods, boosting methods and neural network methods were used. In addition, the prediction and detection of events generated by cybercriminals. The answer is detailed in the Table 5.

3.1.4. Q3: what evaluation metrics have been used to measure the performance of methods for forecasting the rate of cybercrime?

Metrics are key tools to quantitatively evaluate the performance of the methods used. This contributes to the selection, optimization, and validation of cybercrime rate forecasting methods. The main metrics used that were found in the analyzed articles correspond to the classification and regression metrics. The answer is detailed in the Table 6.

Table 4. Results corresponding to Q1

Keyword	Input
Cybercrime prediction	Models for the prediction of cybercrime events and rates in society and the financial sector using various machine learning methods and neural networks. Supervised [1], [13], [27]-[29], [36], [37] and unsupervised [5] machine learning and neural networks [2], [38] are used. It comes across a variety of threat prediction and mitigation models that use diverse datasets and algorithms to identify patterns of cybercrime.
Cyberattack prediction	Models that analyze the prediction of cyberattack such as phishing on websites, the internet and IoT, detection of ransomware and the prediction of malware using machine learning methods and neural networks. It uses supervised [47] machine learning with assembly methods [30], [33]-[35], [39], [40], [41], [43], [48] and the use of neural networks [31], [44]-[46]. For the prediction of cyberattacks, these machine learning and deep learning techniques stand out, which improve efficiency and accuracy.
Cyberfraud detection	Models for the detection of cyberfraud generated by fraudulent transactions mainly in the financial sector with the use of machine learning with assembly methods [26], [30], [32], [51]. These models are used to detect and forecast cyberfraud that use machine learning and deep learning techniques to improve accuracy and efficiency.
Cyberthreat forecasting	Efficient hybrid machine learning (EHML) model with the application of assembly methods for the prediction of cyberthreat [49]. This hybrid model is a different and innovative proposal in the prediction of cyber threats.
Cyberbullying prediction	Models for the prediction of cyberbullying in social networks with the use of machine learning, assembly methods with text data from the social network X (before Twitter) [42] and with malicious images with the application of neural networks [50]. These models assembled with neural network methods for predicting cyberbullying suggest, due to their potential, deepening their application in future work.

Table 5. Results corresponding to Q2

Keyword	Input
Machine learning methods	Simple machine learning methods used with other complex methods in models for predicting cybercrime. Decision tree (DT), KNN, SVM, logistic regression (LR), Naive Bayes (NB), J48 among others were used [1], [2], [13], [27]-[31], [33]-[40], [42], [43], [45], [47]-[51]; and unsupervised methods such as K-means and Gaussian mixture model (GMM) [5]. SVM and LR stand out in their use. These traditional machine learning methods are the most used, but not necessarily the most effective for forecasting cybercrime rates.
Bagging methods	Simple methods that are used in parallel to reduce the variance of estimates in cybercrime prediction models, mainly RF was found [1], [2], [29]-[31], [33]-[37], [39], [40], [42], [43], [45], [49], [51] that stands out in its application, logistic model tree [43], extra tree [36] and enhanced decision tree [49]. Bagging methods are found to be increasingly used to forecast cybercrime rates due to their ability to improve the accuracy of predictions by combining various models.
Boosting methods	Simple methods that are used sequentially to improve the performance of the cybercrime prediction model were eXtreme gradient boosting (XGBoost) [1], [2], [27], [28], [32], [40], LightGBM [1], [26], [40], [44], gradient boosting [1], [2], [41], [51], gradient tree boosting [33], CatBoost [1] and AdaBoost [2]. These boosting methods are relevant for forecasting cybercrime rates due to their ability to manage complex data structures and improve predictive accuracy.
Neural network methods	Neural network methods were found in cybercrime and cyberattack prediction models such as convolutional neural network (CNN) [44], [46], [50], multilayer perceptron (MLP) [2], [38], deep neural network (DNN) [45], and artificial neural network (ANN) [31]. It is found that neural network methods are becoming increasingly relevant for forecasting cybercrime rates, as they take advantage of their ability to process and analyze large amounts of data to predict potential threats. It is suggested that its application be deepened in future work.

Table 6. Results corresponding to Q3

Keyword	Input
Classification metrics	The evaluation metrics accuracy, precision, recall and F1 Score were used to evaluate prediction and detection models for cyberattack [33]-[35], [39]-[41], [43], [45]-[47], cybercrime [2], [5], [13], [27], [28], [29], [36], [37], cyberfraud [26], [30], [32], cyberbullying [42], [50] and cyberthreats [49]. receiver operating characteristic (ROC) curve and area under the curve (AUC) were also applied for the evaluation of cyberattack prediction models [44], [48], cybercrime detection [38] and cyberfraud detection [51]. Studies find that these classification metrics are frequently used to evaluate the accuracy and reliability of the models used in the detection and prediction of cybercrime.
Regression metrics	R-squared (R2) and mean square error (MSE) evaluation metrics were used for cybercrime [1] and cyberattack prediction [31]. These regression metrics that are used to evaluate the performance of forecasting models are mostly used in studies that aim to predict continuous quantities.

3.1.5. Q4: what improvements have been incorporated into the models to forecast the rate of cybercrime?

The improvements that have been incorporated into the models through different methods and techniques seek to optimize the forecast of events and the cybercrime rate. These techniques are used to handle imbalanced data, feature scaling techniques, dimensionality reduction techniques, hyperparameter optimization and the implementation of fixed models. The answer is detailed in the Table 7.

Table 7. Results corresponding to Q4

Keyword	Input
Techniques to handle imbalanced data	Unbalanced datasets caused by scarcity, sampling bias, or recurrent changes in the changing cybercrime context disadvantage data classes and negatively impact evaluation metrics in cybercrime models. Undersampling techniques were applied with the application of Random Undersampling for the detection of cyberfraud [26] and Oversampling with the application of synthetic minority oversampling technique (SMOTE) for the detection of phishing [46], malware [51], fraudulent credit card transactions [32], [51], cybercrime [2] and cyberattack [34]. These techniques were effectively applied for cases of minority classes such as cyberattacks with the disadvantage of a possible decrease in detection rates by reducing false positives. Careful evaluation and adjustment is recommended to ensure optimal performance.
Feature scaling techniques	Feature scaling techniques are applied through data standardization or normalization to improve performance and to make the cybercrime prediction and detection model more effective and precise. The standard scaler [13], [27], [28], [30] and min max scaler [29], [31], [33] techniques are used to bring the features to a similar magnitude, making them comparable and preventing any feature from dominating the algorithm because it has a larger scale. Feature scaling should be only one part of the feature engineering process. Studies show that coding and feature selection are key to improving model performance. In addition, the choice of scaling technique (normalization or standardization) should be aligned with the specific requirements of the machine learning algorithm used.
Dimensionality reduction techniques	Dimensionality reduction techniques are used to eliminate features from data that reduce performance and achieve optimal and efficient predictive models of cybercrime. The principal component analysis (PCA) [1], [32], [39] technique is used to reduce the dimensions of the datasets in the analysis of cyberfraud for fraudulent credit card transactions. These studies show that these techniques simplify large data sets by making them more manageable and interpretable. However, care must be taken in choosing the appropriate technique, depending on the nature of the dataset, to ensure that critical information is not lost.
Hyperparameter optimization	Hyperparameter optimization benefits cybercrime rate forecasting. To achieve the best results in the generation of prediction models, the optimization of the hyperparameters of the prediction models of cyberattack [31], [34], [46], [47], cyberfraud [30], [32], [51], cybercrime [2], [13] and cyberbullying [50] was applied. However, for the optimization of hyperparameters it is necessary to consider the number of computational resources, and the complexity involved in the optimization process.
Mixed models	Mixed models offer significant benefits in cybercrime rate forecasting by integrating multiple analytical techniques to improve prediction accuracy. Combining different methods to improve performance and precision through assembled models for the prediction of cybercrime [1], [2], [13], [27]-[29], [36]-[38], cyberattack [31], [33]-[35], [39]-[41], [43]-[45], [47], cyberfraud [26], [30], [32] and cyberbullying [42]. In addition, hybrid models for the detection of cyberfraud [51], cyberattack [48], cyberthreat [49] and cyberbullying [50]. Studies agree that the implementation of mixed models requires a considerable number of computational resources and the availability of large data sets.

3.2. Bibliometrix analysis

Bibliometric analysis is the quantitative method used in this systematic review. R Studio Bibliometrix software was used to systematically evaluate and visualize the scientific literature regarding trends, patterns, and studies in the field of cybercrime rate forecasting [52], [53]. The VOSViewer tool was also used to build and visualize bibliometric networks such as the cooccurrence network of keywords of the articles and the map of collaboration of studies between countries [54]. A bibliometric analysis of the 229 articles obtained from the search in the databases used in this systematic review was performed.

Figure 4 shows the keyword cloud of the studies in which the word "cybercrime" is most frequently found, followed by the words "detection", "network security", "machine learning", "learning systems", among others. Figure 5 shows the cooccurrence of keywords by color coded periods, revealing that in recent years (in yellow) studies have focused on "cybercrime," "deep learning," and different types of cybercrime and machine learning techniques. Figure 6 shows the trend of publications related to the study in different countries, placing India with the highest number of published articles, followed by the United Kingdom and the United States.

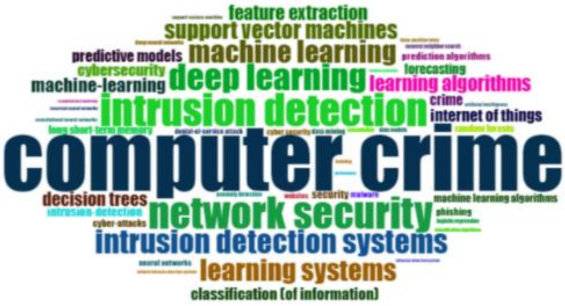

Figure 4. Keyword cloud

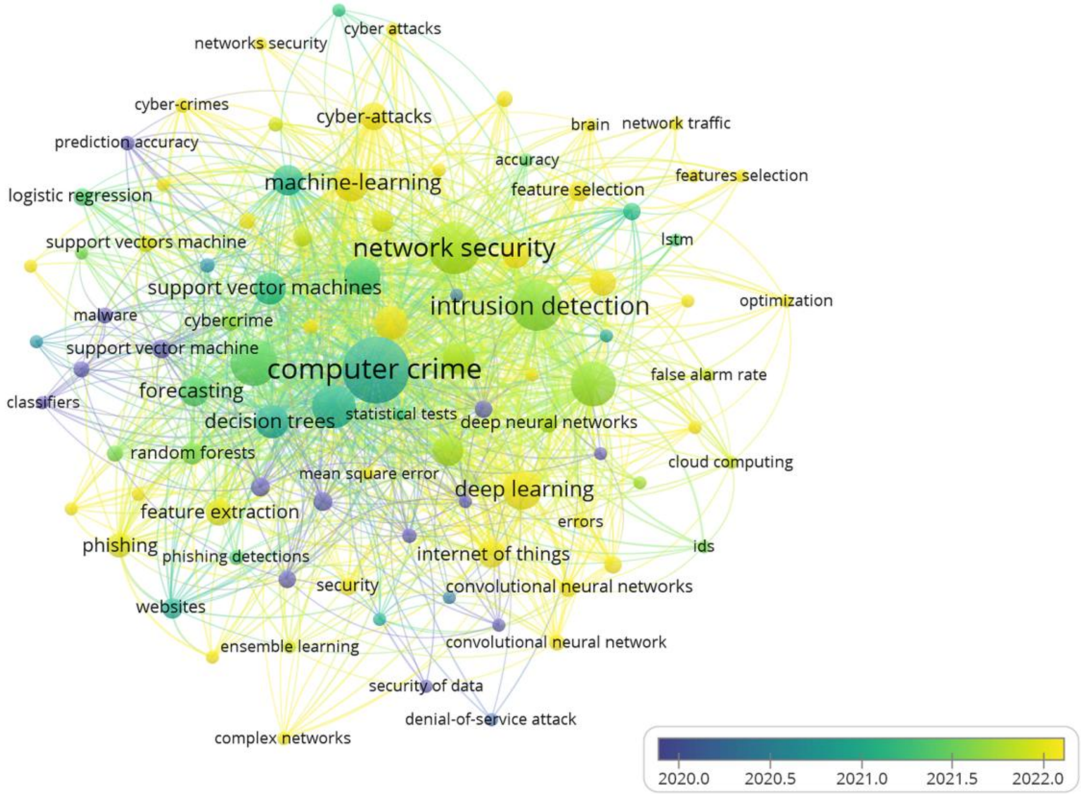

Figure 5. Network to cooccurrence of keywords

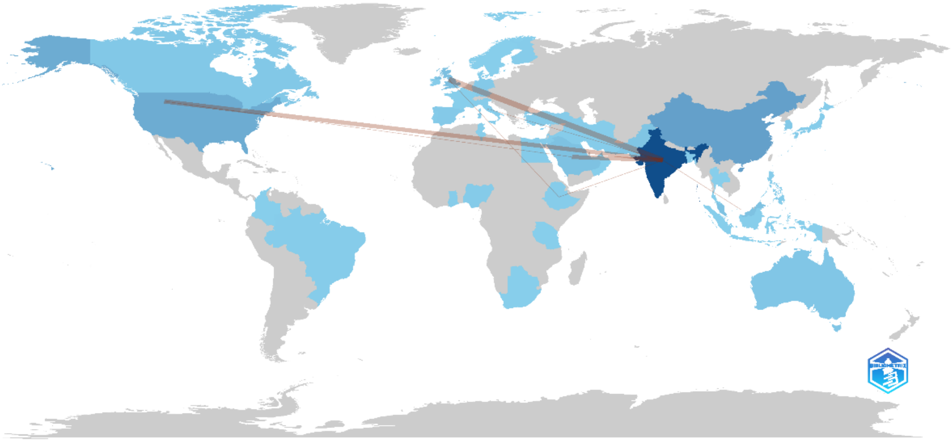

Figure 6. Country collaboration map

3.3. Proposed model

Figure 7 shows the proposed model based on the opportunities for improvement and the good practices found in the review of the studies of this research. The proposal considers the techniques, methods and experiences to build a simple, effective and efficient model to improve the precision of the cybercrime rate forecast.

Cybercrime data series is the historical dataset in time series that includes records of the different cybercrimes. In Data preprocessing, data cleansing is performed to ensure the quality of the actual dataset and its transformation into the data types that the model will process optimally. In Features, techniques are used to balance the unbalanced data in the dataset, then feature reduction techniques for data standardization, and finally techniques to reduce the dimensionality of the data and obtain a tighter predictive model. Data splitting, data is divided into datasets of 70% for training and 30% for testing. In methods, autoregressive integrated moving average (ARIMA) methods would be used to capture the linear and seasonal relationships of a time series and long short term memory (LSTM) to model nonlinear patterns and long term dependencies. The hyperparameter adjustment of the methods would be carried out with optimization algorithms. The results are then assembled to deliver the final forecast.

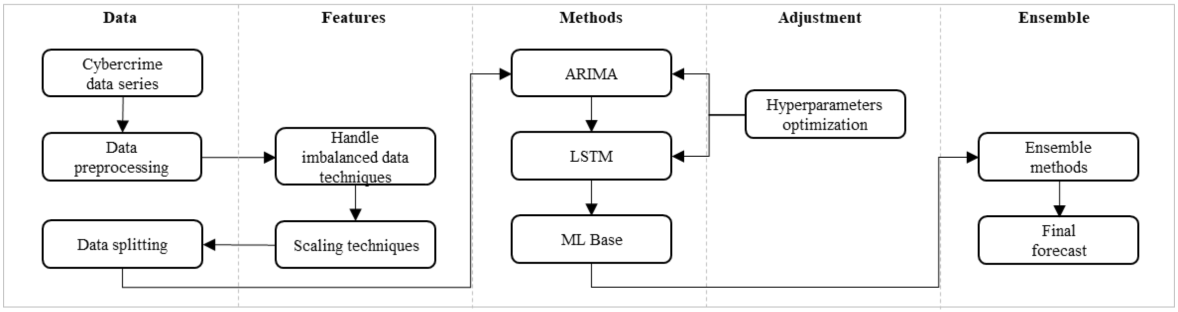

Figure 7. Proposed model

3.4. Discussion

This study conducted the state of the art on the use of artificial intelligence to forecast the rate of cybercrime. The studies [1], [36] address the prediction of cases of common crimes and cybercrimes with machine learning techniques in several cities in India. Another study [44] on Cyber Threats analyzes the prediction of the rate of infectious malware on computers. In the cyberattack, [40] behavioral analysis to improve the prediction rate in phishing detection. Other studies [2], [32], [51] analyze cyberfraud with machine learning models for the detection of bank and credit card fraud. In addition, [42], [50] the prediction of cyberbullying with neural networks for the analysis of messages and images from social networks. While previous studies investigated the use of artificial intelligence to forecast cybercrimes, these are not the most efficient considering new methods combined with neural networks that manage complex data structures that improve predictive accuracy.

Historical datasets are relevant to ensure the accuracy of the cybercrime forecasting model with machine learning. The study [1] used six datasets of common and cybercrime in India collected by the National Crime Records Bureau (NCRB). Another study [49] used Kaggle's "Cybercrime dataset India" with records of cyberattacks across India to forecast crime rates. In addition, [47] used Kaggle's dataset with over 500 thousand URLs for phishing website detection. While these studies propose the use of historical data, other authors improve the efficiency of the model with the use of real-time data such as [29] for the prediction of cyberattacks, with the use of real-time dataset cluster computing techniques for the identification of cybercriminals and [5], [36], [38], [42], they use API connection with the social networks Facebook and X (formerly Twitter) for the extraction of messages from the social network in real time for the prediction of cyberbullying.

The studies apply and compare different machine learning methods to analyze the performance of forecasting cybercrime rates. The study [1] uses several techniques to forecast the rate and timing of digital crimes in cities in India by comparing the performance of six algorithms such as Gradient Boosting, DT, CatBoost, RF, XGBoost and LightGBM using the R2 and MSE evaluation metrics where DT had the highest R2 of 99.9 and the lowest MSE of 0.01. Another study [2] for the detection of cybercrime in the banking sector, it compared the performance of eleven algorithms such as KNN, RF, NB, Gradient Boosting, MLP, DT, AdaBoost, SVM, Linear SVM, Voiting Classifier and XGBoost with evaluation metrics such as accuracy, precision, recall and F1 score where RF stood out with an accuracy of 99.99%. In the prediction of cyberattack, in [29] three algorithms were used, such as LR, RF, and KNN, where the latter had the best performance with an accuracy of 98.78%. Furthermore, [37] compares eight machine learning techniques to predict cyberattacks such as LR, KNN, SVM Linear, SVM Kernel, NB, DT, RF, and XGBoost, where SVM Linear achieved the highest accuracy rate with 66.81% accuracy. While previous studies apply traditional machine learning techniques, they do not address the combination of these techniques or hybrid models to improve forecast performance and accuracy.

Finally, optimizing cybercrime forecasting models through innovative techniques used to achieve the best results. The study [1] for the cybercrime prediction model, it compares the result of six machine learning algorithms with the application of the Standard, Min-Max and PCA techniques for feature scaling or data normalization to find the most effective and accurate model, where in all scenarios the DT algorithm generated the best results in the R2 score and MSE. Another study [32] compares the results of the original dataset against the improved dataset after applying SMOTE, with the latter being the best. For the study [46] the highlight is the reduction of the imbalance in the dataset with the use of the SMOTE technique and the adjustment of hyperparameters for model training, which improved the accuracy of the cyberattack prediction. However, we discovered that [51] in the hybrid model to lower the rate of cyberfraud the SMOTE technique did not achieve a significant improvement in test results.

Our research offers a broad overview of the contribution of artificial intelligence to forecasting the rate of cybercrime. However, we recognize some limitations caused by the scope of studies on the modality of cybercrime that mainly affects people, leaving out prediction models proposed in the field of companies and businesses that will require future research to know the advances of cybersecurity in the prediction of cyberattacks and cyberthreats on this front. Future research can look at these modalities and machine learning methods for forecasting cybercrime in business sectors to confirm and expand on our findings.

In summary, recent advances in the use of artificial intelligence to forecast the rate of cybercrime are growing and innovative. Our findings highlight the potential of machine learning to forecast cybercrime rates by leveraging historical data and traditional and advanced algorithms to predict future trends and identify potential threats. The integration of diverse datasets and the continuous refinement of algorithms are relevant to improving the accuracy and reliability of cybercrime forecasting models that we are sure contribute significantly to the fight against cybercrime.

4. CONCLUSION

After the analysis of the articles included in this systematic review of the literature, the questions posed in this research were answered. It is determined that there are several machine learning models and techniques that have been developed to predict cybercrime rates. Different techniques are used to forecast the rate and timing of cybercrimes, including SVM, LR, RF, XGBoost, MLP, and CNN to predict the types of cyberattacks. The importance of using real data to improve the accuracy of predictions about cybercrime is also highlighted. In addition, the importance of optimizing cybercrime forecasting models through techniques to achieve the best results. Finally, this research provides evidence that could be used effectively by authorities and police forces to formulate strategies, prevention measures, and control of cybercrime for the benefit of society. This research suggests that while significant advances have been made in forecasting the cybercrime rate, new combined or hybrid predictive models such as the ARIMA-LSTM model proposed in this study need to be refined or created to improve the performance and accuracy of cybercrime forecasting.

FUNDING INFORMATION

The authors thank the Universidad Nacional Mayor de San Marcos, Peru. This research contributes to the development of the author's thesis as a requirement for obtaining a doctoral degree and therefore did not receive funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Manuel Martin Morales Barrenechea	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Miguel Angel Cano Lengua		✓		✓	✓			✓		✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] G. Bhardwaj and Dr. R. K. Bawa, "Machine learning techniques based exploration of various types of crimes in India," *Indian Journal of Computer Science and Engineering*, vol. 13, no. 4, pp. 1293–1307, Aug. 2022, doi: 10.21817/indjcse/2022/v13i4/221304142.

[2] A. G. Mohamed, A. Elsayed, and A. A. Galal, "Machine learning for detecting cybercrime in the banking sector," *Journal of Southwest Jiaotong University*, vol. 58, no. 5, pp. 786–799, Oct. 2023, doi: 10.35741/issn.0258-2724.58.5.60.

[3] A. Ampountolas, T. N. Nde, P. Date, and C. Constantinescu, "A Machine Learning Approach for Micro-Credit Scoring," *Risks*, vol. 9, no. 3, p. 50, Mar. 2021, doi: 10.3390/risks9030050.

[4] R. K. Mishra, A. R. Ansari, J. A. A. Jothi, and V. Mishra, "Analysis of Criminal Landscape by Utilizing Statistical Analysis and Deep Learning Techniques," *Journal of Applied Security Research*, vol. 1, pp. 1–26, Feb. 2024, doi: 10.1080/19361610.2024.2314392.

[5] K. Veena, K. Meena, R. Kuppusamy, Y. Teekaraman, R. V. Angadi, and A. R. Thelkar, "Cybercrime: Identification and Prediction Using Machine Learning Techniques," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–10, Aug. 2022, doi: 10.1155/2022/8237421.

[6] S. Goel, "National Cyber Security Strategy and the Emergence of Strong Digital Borders," *Connections: The Quarterly Journal*, vol. 19, no. 1, pp. 73–86, 2020, doi: 10.11610/Connections.19.1.07.

[7] S. Morgan, "Cybercrime To Cost The World $10.5 Trillion Annually By 2025," Cybercrime Magazine. [Online]. Available: https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/. (Accessed: Jul. 16, 2024).

[8] N. AllahRakha, "Impacts of Cybercrimes on the Digital Economy," *Uzbek Journal of Law and Digital Policy*, vol. 2, no. 3, pp. 29–36, Aug. 2024, doi: 10.59022/ujldp.207.

[9] D. Taman, "Impacts of Financial Cybercrime on Institutions and Companies," *Arab Journal of Literature and Humanities*, vol. 8, no. 30, pp. 477–488, Feb. 2024, doi: 10.21608/ajahs.2024.341707.

[10] D. Wright and R. Kumar, "Assessing the socio-economic impacts of cybercrime," *Societal Impacts*, vol. 1, no. 1–2, pp. 1–4, Dec. 2023, doi: 10.1016/j.socimp.2023.100013.

[11] A. D. Sharko, G. Sharko, and S. Qose, "Artificial Intelligence In Cybersecurity Applications," in *2024 IEEE 28th International Conference on Intelligent Engineering Systems (INES)*, IEEE, Jul. 2024, pp. 175–180, doi: 10.1109/INES63318.2024.10629129.

[12] A. Parisi, *Hands-on artificial intelligence for cybersecurity: implement smart AI systems for preventing cyber attacks and detecting threats and network anomalies*, 1st ed., vol. 1. Birmingham: Packt Publishing, 2019.

[13] E. F. Aljarboua, M. Bte Md. Din, and A. A. Bakar, "Cyber-Crime Detection: Experimental Techniques Comparison Analysis," in *2022 International Visualization, Informatics and Technology Conference (IVIT)*, IEEE, Nov. 2022, pp. 124–129, doi: 10.1109/IVIT55443.2022.10033332.

[14] K. Veena, K. Meena, Y. Teekaraman, R. Kuppusamy, and A. Radhakrishnan, "SVM Classification and KNN Techniques for Cyber Crime Detection," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–9, Jan. 2022, doi: 10.1155/2022/3640017.

[15] P. Boonyopakorn, N. Wisitpongphan, and U. Changsan, "Classifying Cybercrime and Threat on Thai Online News: A Comparison of Supervised Learning Algorithms," in *2023 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)*, IEEE, Jun. 2023, pp. 1–6, doi: 10.1109/ITC-CSCC58803.2023.10212562.

[16] X. Fang, M. Xu, S. Xu, and P. Zhao, "A deep learning framework for predicting cyber attacks rates," *EURASIP Journal on Information Security*, vol. 2019, no. 1, p. 5, Dec. 2019, doi: 10.1186/s13635-019-0090-6.

[17] R. Ch, T. R. Gadekallu, M. H. Abidi, and A. Al-Ahmari, "Computational System to Classify Cyber Crime Offenses using Machine Learning," *Sustainability*, vol. 12, no. 10, p. 4087, May 2020, doi: 10.3390/su12104087.

[18] D. M. Cao *et al.*, "Advanced Cybercrime Detection: A Comprehensive Study on Supervised and Unsupervised Machine Learning Approaches Using Real-world Datasets," *Journal of Computer Science and Technology Studies*, vol. 6, no. 1, pp. 40–48, Jan. 2024, doi: 10.32996/jcsts.2024.6.1.5.

[19] A. Falade, A. Azeta, A. Oni, and I. Odun-ayo, "Systematic Literature Review of Crime Prediction and Data Mining," *Review of Computer Engineering Studies*, vol. 6, no. 3, pp. 56–63, Nov. 2019, doi: 10.18280/rces.060302.

[20] R. van Dinter, B. Tekinerdogan, and C. Catal, "Automation of systematic literature reviews: A systematic literature review," *Information and Software Technology*, vol. 136, p. 106589, Aug. 2021, doi: 10.1016/j.infsof.2021.106589.

[21] Y. Harie, B. Gautam, and K. Wasaki, "Computer Vision Techniques for Growth Prediction: A Prisma-Based Systematic Literature Review," *Applied Sciences*, vol. 13, p. 5335, Oct. 2023, doi: 10.3390/app13095335.

[22] A. Kumar, "Systematic Literature Review (SLR)," in *Meta-analysis in Clinical Research: Principles and Procedures*, vol. 1, 2023, pp. 7–14, doi: 10.1007/978-981-99-2370-0_2.

[23] M. S. Cumpston, J. E. McKenzie, R. Ryan, E. Flemyng, J. Thomas, and S. E. Brennan, "Development of the InSynQ checklist: A tool for planning and reporting the synthesis questions in systematic reviews of interventions," *Cochrane Evidence Synthesis and Methods*, vol. 1, no. 10, pp. 1–13, Dec. 2023, doi: 10.1002/cesm.12036.

[24] B. Shui, Z. Cai, and X. Luo, "Towards customized mitigation strategy in the transportation sector: An integrated analysis framework combining LMDI and hierarchical clustering method," *Sustainable Cities and Society*, vol. 107, p. 105340, Jul. 2024, doi: 10.1016/j.scs.2024.105340.

[25] J. R. N. Villar and M. A. C. Lengua, "A Systematic Review of the Literature on the Use of Artificial Intelligence in Forecasting the Demand for Products and Services in Various Sectors," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, pp. 144–156, 2024, doi: 10.14569/IJACSA.2024.0150315.

[26] M. Guan, R. Xue, Z. Wu, H. Yang, D. Song, and Z. Zhang, "A high performance fraud detection strategy prediction model," in *2022 2nd International Conference on Computer Science and Blockchain (CCSB),* IEEE, Oct. 2022, pp. 107–110, doi: 10.1109/CCSB58128.2022.00026.

[27] C. Singh, R. Singh, Shivaputra, M. Tiwari, and B. Hazela, "Analyse and Predict the Detection of the Cyber - Attack Process by Using a Machine-Learning Approach," *EAI Endorsed Transactions on Internet of Things*, vol. 10, pp. 1–6, Mar. 2024, doi: 10.4108/eetiot.5345.

[28] A. Bilen and A. B. Özer, "Cyber-attack method and perpetrator prediction using machine learning algorithms," *PeerJ Computer Science*, vol. 7, pp. 1–21, Apr. 2021, doi: 10.7717/peerj-cs.475.

[29] A. Swaminathan, B. Ramakrishnan, K. M, and S. R, "Prediction of Cyber-attacks and Criminality Using Machine Learning Algorithms," in *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT),* IEEE, Nov. 2022, pp. 547–552, doi: 10.1109/3ICT56508.2022.9990652.

[30] N. Al-Ghamdi and T. Alsubait, "Digital Forensics and Machine Learning to Fraudulent Email Prediction," in *2022 Fifth National Conference of Saudi Computers Colleges (NCCC)*, IEEE, Dec. 2022, pp. 99–106, doi: 10.1109/NCCC57165.2022.10067685.

[31] R. Verma and B. Thakur, "Machine Learning Techniques for the Prediction of Cyber-Attacks," in *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, IEEE, Nov. 2023, pp. 978–985, doi: 10.1109/ICCCIS60361.2023.10425542.

[32] P. A. Jadhav, U. Lalwani, A. Gour, M. Shayan, and S. Motwani, "Identification of Fraudulent Credit Card transactions using Machine Learning Algorithms," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, IEEE, Apr. 2022, pp. 1–4, doi: 10.1109/I2CT54291.2022.9824165.

[33] K. Amen, M. Zohdy, and M. Mahmoud, "Machine Learning for Multiple Stage Phishing URL Prediction," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, Dec. 2021, pp. 794–800, doi: 10.1109/CSCI54926.2021.00049.

[34] A. Sharma and H. Babbar, "Machine Learning Solutions for Evolving Injection Attack Landscape," in *2023 2nd International Conference on Futuristic Technologies (INCOFT)*, IEEE, Nov. 2023, pp. 1–6, doi: 10.1109/INCOFT60753.2023.10425456.

[35] A. O. Almashhadani, M. Kaiiali, S. Sezer, and P. O'Kane, "A Multi-Classifier Network-Based Crypto Ransomware Detection System: A Case Study of Locky Ransomware," *IEEE Access*, vol. 7, pp. 47053–47067, 2019, doi: 10.1109/ACCESS.2019.2907485.

[36] D. D. Pandya, G. Amarawat, A. Jadeja, S. Degadwala, and D. Vyas, "Analysis and Prediction of Location based Criminal Behaviors Through Machine Learning," in *2022 International Conference on Edge Computing and Applications (ICECAA)*, IEEE, Oct. 2022, pp. 1324–1332, doi: 10.1109/ICECAA55415.2022.9936498.

[37] N. S. Deepak, T. Hanitha, K. Tanniru, L. R. Kiran, N. R. Sai, and M. J. Kumar, "Analyze and Forecast the Cyber Attack Detection Process using Machine Learning Techniques," in *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, Jul. 2023, pp. 1732–1738, doi: 10.1109/ICESC57686.2023.10193289.

[38] T. Arora, M. Sharma, and S. K. Khatri, "Detection of Cyber Crime on Social Media using Random Forest Algorithm," in *2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC)*, IEEE, Oct. 2019, pp. 47–51, doi: 10.1109/PEEIC47157.2019.8976474.

[39] S. Naaz, "Detection of Phishing in Internet of Things Using Machine Learning Approach," *International Journal of Digital Crime and Forensics*, vol. 13, no. 2, pp. 1–15, Mar. 2021, doi: 10.4018/IJDCF.2021030101.

[40] A. R. Omar, S. A. Taie, and M. E.Shaheen, "From Phishing Behavior Analysis and Feature Selection to Enhance Prediction Rate in Phishing Detection," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 1033–1044, 2023, doi: 10.14569/IJACSA.2023.01405107.

[41] J. Philomina, K. A. F. Fathima, S. Gayathri, G. E. Elias, and A. A. Menon, "A comparitative study of machine learning models for the detection of Phishing Websites," in *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, IEEE, Jun. 2022, pp. 1–7, doi: 10.1109/IC3SIS54991.2022.9885595.

[42] S. M. M. Matias, J. A. Costales, and C. M. De Los Santos, "A Framework for Cybercrime Prediction on Twitter Tweets Using Text-Based Machine Learning Algorithm," in *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, IEEE, Aug. 2022, pp. 235–240, doi: 10.1109/PRAI55851.2022.9904212.

[43] R. M. A. Latif, M. Umer, T. Tariq, M. Farhan, O. Rizwan, and G. Ali, "A Smart Methodology for Analyzing Secure E-Banking and E-Commerce Websites," in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, IEEE, Jan. 2019, pp. 589–596, doi: 10.1109/IBCAST.2019.8667255.

[44] A. bin Asad, R. Mansur, S. Zawad, N. Evan, and M. I. Hossain, "Analysis of Malware Prediction Based on Infection Rate Using Machine Learning Techniques," in *2020 IEEE Region 10 Symposium (TENSYMP)*, IEEE, 2020, pp. 706–709, doi: 10.1109/TENSYMP50017.2020.9230624.

[45] P. T. Devadarshini, B. Chandrashekar, S. Pundir, M. Tiwari, R. Madala, and E. Indhuma, "Cognitive Defense Cyber Attack Prediction and Security Design in Machine Learning Model," in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, Sep. 2023, pp. 1361–1366, doi: 10.1109/IC3I59117.2023.10397602.

[46] P. Kotian and R. Sonkusare, "Detection of Malware in Cloud Environment using Deep Neural Network," in *2021 6th International Conference for Convergence in Technology (I2CT)*, IEEE, Apr. 2021, pp. 1–5, doi: 10.1109/I2CT51068.2021.9417901.

[47] N. Varsha, D. P. Kumar, B. Akhil, and K. Mouli, "Phishing sites detection using Machine Learning," in *2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, IEEE, Nov. 2023, pp. 1–7, doi: 10.1109/ICIICS59993.2023.10421336.

[48] A. O. Balogun, K. S. Adewole, A. O. Bajeh, and R. G. Jimoh, "Cascade Generalization Based Functional Tree for Website Phishing Detection," in *Communications in Computer and Information Science*, vol. 1487 CCIS, Springer Science and Business Media Deutschland GmbH, 2021, pp. 288–306, doi: 10.1007/978-981-16-8059-5_17.

[49] U. K. Lilhore, S. Simaiya, J. K. Sandhu, A. Baliyan, and A. Garg, "EHML: An Efficient Hybrid Machine Learning Model for Cyber Threat Forecasting in CPS," in *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, IEEE, Jan. 2023, pp. 1453–1458, doi: 10.1109/AISC56616.2023.10084987.

[50] M. Elmezain, A. Malki, I. Gad, and E.-S. Atlam, "Hybrid deep learning model-based prediction of images related to cyberbullying," *International Journal of Applied Mathematics and Computer Science*, vol. 32, no. 2, pp. 323–334, Jun. 2022, doi: 10.34768/amcs-2022-0024.

[51] D. Sharma and S. S.Kang, "Hybrid model for detection of frauds in credit cards," in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, IEEE, Dec. 2022, pp. 70–77, doi: 10.1109/ICAC3N56670.2022.10074057.

[52] N. Senthilvadevel *et al.*, “Evaluating global research trends in special needs dentistry: A systematic bibliometrix analysis,” *Clin Exp Dent Res*, vol. 10, no. 3, p. 1, Jun. 2024, doi: 10.1002/cre2.896.

[53] A. Artyukhov, A. Lapidus, O. Yeremenko, N. Artyukhova, and O. Churikanova, “An R Studio Bibliometrix Analysis of Global Research Trends of Educational Crises in 2020s,” *SocioEconomic Challenges*, vol. 8, no. 2, pp. 88–108, Jul. 2024, doi: 10.61093/sec.8(2).88-108.2024.

[54] H. El Bekkouri *et al.*, “Bibliometric Analysis of the Literature on Carbon Ion Therapy Using VOSviewer Software and Dimensions Database,” *Atom Indonesia*, vol. 50, no. 2, pp. 183–189, Aug. 2024, doi: 10.55981/aij.2024.1392.

BIOGRAPHIES OF AUTHORS

Manuel Martin Morales Barrenechea is a professor at Universidad Peruana de Ciencias Aplicadas (UPC), a Systems Engineer from Universidad Peruana de Ciencias Aplicadas, a Master's in Business Administration and Management from Universidad Alas Peruanas (UAP) and a PhD candidate in Systems and Computer Engineering from Universidad Nacional Mayor de San Marcos (UNMSM). He is Head of Digital Innovation Projects at a leading company in the Telecommunications sector. He can be contacted at email: martin.moralesb@unmsm.edu.pe.

Miguel Angel Cano Lengua is a professor at Universidad Tecnológica del Perú (UTP) and Universidad Nacional Mayor de San Marcos (UNMSM), has a degree in Mathematics, a Ph.D. Engineering of Systems and Computer Science from Universidad Nacional Mayor de San Marcos, a Master's in Systems Engineering from the Universidad Nacional del Callao (UNAC). He works on continuous optimization, artificial intelligence algorithms, conical programming, numerical methods, methodology, and software design. He can be contacted at email: mcanol@unmsm.edu.pe.