НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
НАВЧАЛЬНО-НАУКОВИЙ КОМПЛЕКС
«ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ»
НАЦІОНАЛЬНОГО ТЕХНІЧНОГО УНІВЕРСИТЕТУ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

СИСТЕМНІ ДОСЛІДЖЕННЯ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

МІЖНАРОДНИЙ НАУКОВО–ТЕХНІЧНИЙ ЖУРНАЛ

№ 4 2023

ЗАСНОВАНО У ЛИПНІ 2001 р.

АДРЕСА РЕДАКЦІЇ:
03056, м. Київ,
просп. Перемоги, 37, корп. 35,
ННК «ІПСА» КПІ ім. Ігоря Сікорського
Тел.: **204-81-44**; факс: **204-81-44**
E-mail: journal.iasa@gmail.com
http://journal.iasa.kpi.ua

У номері:

• Проблеми прийняття рішень та управління в економічних, технічних, екологічних і соціальних системах

• Теоретичні та прикладні проблеми інтелектуальних систем підтримання прийняття рішень

• Проблемно і функціонально орієнтовані комп'ютерні системи та мережі

• Евристичні методи та алгоритми в системному аналізі та управлінні

NATIONAL ACADEMY OF SCIENCES OF UKRAINE
EDUCATIONAL AND SCIENTIFIC COMPLEX
«INSTITUTE FOR APPLIED SYSTEM ANALYSIS»
OF THE NATIONAL TECHNICAL UNIVERSITY OF UKRAINE
«IGOR SIKORSKY KYIV POLYTECHNIC INSTITUTE»

SYSTEM RESEARCH AND INFORMATION TECHNOLOGIES

INTERNATIONAL SCIENTIFIC AND TECHNICAL JOURNAL

№ 4 2023

IT IS FOUNDED IN JULY 2001

THE EDITION ADDRESS:

03056, Kyiv,
av. Peremogy, 37, building 35,
Institute for Applied System Analysis
at the Igor Sikorsky Kyiv Polytechnic Institute
Phone: **204-81-44**; Fax: **204-81-44**
E-mail: journal.iasa@gmail.com
http://journal.iasa.kpi.ua

In the issue:

- **Decision making and control in economic, technical, ecological and social systems**
- **Theoretical and applied problems of intelligent systems for decision making support**
- **Problem- and function-oriented computer systems and networks**
- **Heuristic methods and algorithms in system analysis and control**

Шановні читачі!

Навчально-науковий комплекс «Інститут прикладного системного аналізу» Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» видає міжнародний науково-технічний журнал

«СИСТЕМНІ ДОСЛІДЖЕННЯ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ».

Журнал публікує праці теоретичного та прикладного характеру в широкому спектрі проблем, що стосуються системних досліджень та інформаційних технологій.

Провідні тематичні розділи журналу:

Теоретичні та прикладні проблеми і методи системного аналізу; теоретичні та прикладні проблеми інформатики; автоматизовані системи управління; прогресивні інформаційні технології, високопродуктивні комп'ютерні системи; проблеми прийняття рішень і управління в економічних, технічних, екологічних і соціальних системах; теоретичні та прикладні проблеми інтелектуальних систем підтримання прийняття рішень; проблемно і функціонально орієнтовані комп'ютерні системи та мережі; методи оптимізації, оптимальне управління і теорія ігор; математичні методи, моделі, проблеми і технології дослідження складних систем; методи аналізу та управління системами в умовах ризику і невизначеності; евристичні методи та алгоритми в системному аналізі та управлінні; нові методи в системному аналізі, інформатиці та теорії прийняття рішень; науково-методичні проблеми в освіті.

Головний редактор журналу — ректор Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», академік НАН України Михайло Захарович Згуровський.

Журнал «Системні дослідження та інформаційні технології» включено до переліку фахових видань ВАК України.

Журнал «Системні дослідження та інформаційні технології» входить до таких наукометричних баз даних: Scopus, EBSCO, Google Scholar, DOAJ, Index Copernicus, реферативна база даних «Україніка наукова», український реферативний журнал «Джерело», наукова періодика України.

Статті публікуються українською та англійською мовами.

Журнал рекомендовано передплатити. **Наш індекс 23918.** Якщо ви не встигли передплатити журнал, його можна придбати безпосередньо в редакції за адресою: 03056, м. Київ, просп. Перемоги, 37, корп. 35.

Завідувачка редакції **С.М. Шевченко**
Редакторка **Р.М. Шульженко**
Молодша редакторка **Л.О. Тарин**
Комп'ютерна верстка, дизайн **А.А. Патіохи**

Свідоцтво про реєстрацію КВ № 23234–13074 ПР від 22.03.2018 р.

Підписано до друку 26.12.2023. Формат 70x108 1/16. Папір офс. Гарнітура Times.
Спосіб друку – цифровий. Ум. друк. арк. 14,411. Обл.-вид. арк. 28,56. Наклад 87 пр. Зам. № 11/04

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Свідоцтво про державну реєстрацію: ДК № 5354 від 25.05.2017 р.
просп. Перемоги, 37, м. Київ, 03056.
ФОП Пилипенко Н.М., вул. Мічуріна, б. 2/7, м. Київ, 01014.
Виписка з Єдиного державного реєстру № 2 070 000 0000 0214697 від 17.05.2019 р.,
тел.. (044) 361 78 68.

Dear Readers!

Educational and Scientific Complex «Institute for Applied System Analysis» of the National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute» is published of the international scientific and technical journal

«SYSTEM RESEARCH AND INFORMATION TECHNOLOGIES».

The Journal is printing works of a theoretical and applied character on a wide spectrum of problems, connected with system researches and information technologies.

The main thematic sections of the Journal are the following:

Theoretical and applied problems and methods of system analysis; theoretical and applied problems of computer science; automated control systems; progressive information technologies, high-efficiency computer systems; decision making and control in economic, technical, ecological and social systems; theoretical and applied problems of intellectual systems for decision making support; problem- and function-oriented computer systems and networks; methods of optimization, optimum control and theory of games; mathematical methods, models, problems and technologies for complex systems research; methods of system analysis and control in conditions of risk and uncertainty; heuristic methods and algorithms in system analysis and control; new methods in system analysis, computer science and theory of decision making; scientific and methodical problems in education.

The editor-in-chief of the Journal is rector of the National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», academician of the NASU Michael Zaharovich Zgurovsky.

The articles to be published in the Journal in Ukrainian and English languages are accepted. Information printed in the Journal is included in the Catalogue of periodicals of Ukraine.

СИСТЕМНІ ДОСЛІДЖЕННЯ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

4 • 2023

ЗМІСТ

ПРОБЛЕМИ ПРИЙНЯТТЯ РІШЕНЬ ТА УПРАВЛІННЯ В ЕКОНОМІЧНИХ, ТЕХНІЧНИХ, ЕКОЛОГІЧНИХ І СОЦІАЛЬНИХ СИСТЕМАХ

Vedna Sharma, Surender Singh Samant. A multi-level decision-making framework for heart-related disease prediction and recommendation 7

Donets V.V., Strilets V.Y., Ugryumov M.L., Shevchenko D.O., Prokopovych S.V., Chagovets L.O. Methodology of the countries' economic development data analysis 21

Dorogyy Ya., Kolisnichenko V. Blockchain transaction analysis: a comprehensive review of applications, tasks and methods 37

ТЕОРЕТИЧНІ ТА ПРИКЛАДНІ ПРОБЛЕМИ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ ПІДТРИМАННЯ ПРИЙНЯТТЯ РІШЕНЬ

Мартьянов Д., Виклюк Я., Флейчук М. Моделювання динаміки ринку криптовалют з використанням інструментів машинного навчання 54

Bazdyrev A. Semi-supervised inverted file index approach for approximate nearest neighbor search 69

ПРОБЛЕМНО І ФУНКЦІОНАЛЬНО ОРІЄНТОВАНІ КОМП'ЮТЕРНІ СИСТЕМИ ТА МЕРЕЖІ

Hend Khalid Alkahtani. Raising the information security awareness among social media users in the Middle East 76

Pradip M. Paithane, Sarita Jibhau Wagh. Novel modified kernel fuzzy c-means algorithm used for cotton leaf spot detection 85

Bazilevych K., Kyrylenko O., Parfeniuk Y., Yakovlev S., Krivtsov S., Meniailov I., Kuznietcova V., Chumachenko D. Information system for assessing the informativeness of an epidemic process features 100

ЕВРИСТИЧНІ МЕТОДИ ТА АЛГОРИТМИ В СИСТЕМНОМУ АНАЛІЗІ ТА УПРАВЛІННІ

Chaudhari S., Aparna R. Survey of image deduplication for cloud storage 113

Suman Sansanwal, Nitin Jain. A comprehensive survey on load balancing techniques for virtual machines 135

Відомості про авторів 148

Зміст журналу за 2023р. 150

Автори статей за 2023р. 152

SYSTEM RESEARCH AND INFORMATION TECHNOLOGIES

4 • 2023

CONTENT

DECISION MAKING AND CONTROL IN ECONOMIC, TECHNICAL, ECOLOGICAL AND SOCIAL SYSTEMS

Vedna Sharma, Surender Singh Samant. A multi-level decision-making framework for heart-related disease prediction and recommendation 7

Donets V.V., Strilets V.Y., Ugryumov M.L., Shevchenko D.O., Prokopovych S.V., Chagovets L.O. Methodology of the countries' economic development data analysis 21

Dorogyy Ya., Kolisnichenko V. Blockchain transaction analysis: a comprehensive review of applications, tasks and methods 37

THEORETICAL AND APPLIED PROBLEMS OF INTELLIGENT SYSTEMS FOR DECISION MAKING SUPPORT

Martjanov D., Vyklyuk Ya., Fleychuk M. Modeling cryptocurrency market dynamics using machine learning tools 54

Bazdyrev A. Semi-supervised inverted file index approach for approximate nearest neighbor search 69

PROBLEM- AND FUNCTION-ORIENTED COMPUTER SYSTEMS AND NETWORKS

Hend Khalid Alkahtani. Raising the information security awareness among social media users in the Middle East 76

Pradip M. Paithane, Sarita Jibhau Wagh. Novel modified kernel fuzzy c-means algorithm used for cotton leaf spot detection 85

Bazilevych K., Kyrylenko O., Parfeniuk Y., Yakovlev S., Krivtsov S., Meniailov I., Kuznietcova V., Chumachenko D. Information system for assessing the informativeness of an epidemic process features 100

HEURISTIC METHODS AND ALGORITHMS IN SYSTEM ANALYSIS AND CONTROL

Chaudhari S., Aparna R. Survey of image deduplication for cloud storage 113

Suman Sansanwal, Nitin Jain. A comprehensive survey on load balancing techniques for virtual machines 135

Information about the authors 148

ПРОБЛЕМИ ПРИЙНЯТТЯ РІШЕНЬ ТА УПРАВЛІННЯ В ЕКОНОМІЧНИХ, ТЕХНІЧНИХ, ЕКОЛОГІЧНИХ І СОЦІАЛЬНИХ СИСТЕМАХ

UDC 62-50
DOI: 10.20535/SRIT.2308-8893.2023.4.01

A MULTI-LEVEL DECISION-MAKING FRAMEWORK FOR HEART-RELATED DISEASE PREDICTION AND RECOMMENDATION

VEDNA SHARMA, SURENDER SINGH SAMANT

Abstract. The precise prediction of health-related issues is a significant challenge in healthcare, with heart-related diseases posing a particularly threatening global health problem. Accurate prediction and recommendation for heart-related diseases are crucial for timely and effective treatment solutions. The primary objective of this study is to develop a classification model capable of accurately identifying heart diseases and providing appropriate recommendations for patients. The proposed system utilizes a multilevel-based classification mechanism employing Support Vector Machines. It aims to categorize heart diseases by analyzing patient's vital parameters. The performance of the proposed model was evaluated by testing it on a dataset containing patient records. The generated recommendations are based on a comprehensive assessment of the severity of clinical features exhibited by patients, including estimating the associated risk of both clinical features and the disease itself. The predictions were evaluated using three metrics: accuracy, specificity, and the receiver operating characteristic curve. The proposed Multilevel Support Vector Machine (MSVM) classification model achieved an accuracy rate of 94.09% in detecting the severity of heart disease. This makes it a valuable tool in the medical field for providing timely diagnosis and treatment recommendations. The proposed model presents a promising approach for accurately predicting heart-related diseases and highlights the potential of soft computing techniques in healthcare. Future research could focus on further enhancing the proposed model's accuracy and applicability.

Keywords: healthcare, heart disease, classification model, learning techniques.

INTRODUCTION

In recent years, a large amount of medical data has become available, representing the healthcare status of patients. This data includes medical reports, test results, and lab reports. Data mining plays a significant role in healthcare recommendation systems, as it enables healthcare professionals to extract valuable insights from the data. These insights can be used to provide more accurate and personalized recommendations to patients [1], [2]. In the current era, people are facing major health issues due to inactive lifestyles. The online healthcare system has proven to be beneficial, especially in scenarios like the COVID-19 pandemic, and has gained significant attention from researchers. Integrating recommender systems into healthcare can support doctors, medical professionals, and patients. These systems assist patients in improving their health conditions and adopting a healthier lifestyle. Healthcare recommendation systems have evolved with various learning technologies and big data science, which provide online suggestions to patients regarding their health issues.

Machine Learning's application in healthcare is poised to revolutionize the industry, as it enhances capabilities and reduces expenses. This progress empowers healthcare practitioners, including doctors and personnel, to focus more on delivering improved patient care. Many researchers have contributed to healthcare recommendation systems for diagnosing various diseases. Previous studies have employed machine learning classifiers and deep learning techniques to predict different diseases using large datasets gathered from various healthcare repositories.

Recommender systems face various challenges, such as reliability, accuracy, dependability, data loss, and issues related to data integrity and quality. To overcome these challenges, various classification models have been proposed, including the usage of different soft computing techniques like machine learning and deep learning. The significance of this research is to enrich the healthcare dataset for better prediction of multidisciplinary diseases. This study proposes an intelligent disease classification mechanism that aims to address various issues in existing systems by predicting the risks associated with diseases. In this study, a disease classification model is used to predict the risks related to heart conditions. The multi-classification methodology works accurately and provides better results in larger healthcare datasets.

The utilization of the fuzzy technique enables the provision of recommendations to patients based on the severity score. The organization of the paper is as follows: Section 2 reviews recent works in the healthcare recommendation field, Section 3 outlines the research methodology employed in this study, and Section 4 presents the dataset and experimental results. Finally, Section 5 concludes the paper and discusses future plans.

RELATED WORK

Several studies have been conducted on healthcare recommender systems, targeting specific diseases, health-related issues, and recommender contexts. These existing studies have highlighted the need for a comprehensive overview supported by a healthcare recommendation system in various recommendation scenarios. Healthcare recommendation systems offer better personalization, increasing user understanding of their medical conditions [4]. These systems are also concerned with providing accurate information, assisting users, and ensuring the security and privacy of patient information. Traditionally, doctors relied on invasive diagnostic methods to identify heart disease, involving an assessment of the physical examination findings, medical history of patients' and investigation into associated symptoms [5].

Cardiovascular disease, including coronary artery disease, is a major global cause of death, particularly among middle-aged and elderly individuals. Traditional diagnostic methods, such as angiography, are expensive and have notable side effects. To explore alternative approaches, researchers have extensively studied data mining techniques and machine learning techniques. In one proposed work for the accurate diagnosis of coronary heart related disease, the performance of a neural network improved by approximately 10% through the application of genetic algorithms to optimize the network's initial weights [6].

Diagnosing and treating heart disease becomes extremely challenging in underdeveloped countries, where there is a lack of necessary medical tools and specialized professionals [7]. Basic and Deep Neural Networks have shown efficiency in exploratory comparative trials, with the deep neural network outperforming most other methods [8].

Subiksha et al. [9] designed a framework for a medical care system based on machine learning. The framework, based on a decentralized network, was designed to link various healthcare databases and services.

Sahoo et al. [10] developed an advanced prediction model. Their study introduces an intelligent health recommendation system (HRS) that leverages big data analytics to implement an efficient health recommendation engine. The proposed intelligent HRS outperforms existing approaches by achieving a lower mean absolute error (MAE) value, transforming the healthcare industry into a more personalized paradigm within a telehealth environment.

Archenna et al. [11] proposed a methodology for generating a healthcare system. They employed big data analytics in the proposed recommendation system, demonstrating how and where to apply big data technologies to construct an efficient patient recommender system. The study emphasized the need for a system capable of handling massive amounts of semi-structured and unstructured patient information, as well as streaming live information about patients from various social media activities. By utilizing the appropriate machine learning (ML) tools and simulations offered by Apache Spark, useful insights can be derived from vast amounts of medical information. The proposed health recommendation system could anticipate a patient's medical condition by assessing their lifestyle, general medical variables, cognitive medical conditions, and interpersonal networks.

Vanisree K. et al. [12] discuss the significance of an early diagnosis of Congenital Heart Disease (CHD) and propose a Decision Support System to improve accuracy and reduce costs. This system was developed using MATLAB's GUI feature and incorporates a Backpropagation Neural Network.

Abugabah et al. [13] designed a medical care analyzer system based on data mining methodology. In this research work, an optimization-based approach based on neural network was implemented to achieve efficient results. Clinical information was retrieved and normalized using a min-max normalization technique. The patient's condition was examined and categorized as either healthy or unhealthy. The supervised learning approach utilized the harmonic optimized modularity neural network.

Mudaliar et al. [14] developed an application programming interface (API) that utilizes the machine learning algorithms to analyze a user's symptoms and diagnose a specific disease. The framework also recommends suitable drugs for users afflicted with that disease. The prediction of illness probability takes into account externally observable symptoms such as temperature, cough, headache, and other indications experienced by a person. The Naive Bayes algorithm was employed to diagnose the illness based on these signs.

In a study conducted by Yoo et al. [8], a medical care recommendation system based on data mining was developed. The proposed system utilized a peer-to-peer collection and adaptable judgment response. Handheld sensors were employed to gather public health records regarding various aspects of an individual's life, including nutrition, daily routines, sleeping patterns, lifestyle behaviors, and occupational stress. Validated index data from the P2P-dataset and personal identification were also considered. A mobile service-based medical care recommendation cellular modem could be designed to enhance the quality of care for patient user-based health management, reduce medical costs, and improve service perception of quality in the medicinal industry. Table 1 provides a summary of related work in healthcare recommendation systems.

Based on the literature review, it can be concluded that selecting important features based on studies such as [16–18; 28–31] and employing a combination of classifiers as demonstrated in [16–18; 20; 28–31] can significantly enhance the predictive capabilities of machine learning algorithms for early-stage detection of heart disease. However, feature selection presents a challenge due to the exponential increase in complexity as the number of features in the dataset grows. Evaluating all possible feature subsets becomes computationally intensive and impractical as the number of features increases [32]. Therefore, alternative strategies are necessary to address this issue.

From Table 1, it is evident that only a limited number of research studies have focused on optimizing hyper parameters due to the time-consuming process of tuning multiple machine-learning models to find the best hyper parameters. However, effectively optimizing classifier hyper parameters can greatly improve the accuracy of predicting the risk of coronary heart disease. Healthcare recommendation systems can benefit from the simplicity and accessibility of the multiclass Support Vector Machine (SVM) model, as it has a limited number of hyper parameters for tuning. This characteristic simplifies the training and optimization process. In contrast to more complex machine learning algorithms, the proposed multi classification-based SVM model's straightforwardness makes it a convenient choice for implementation in healthcare recommendation systems.

Table 1. Summary of Related Work

Author References	Proposed Work	Outcomes & Limitations	Scope for future work
Subiksha, et al., [9] (2018)	The deep learning-based health analyzer system. Performance metrics are Precision and Recall	High error rate	Need to enhance the methodology for better results in information retrieval
Sahoo et al., [10] (2019)	Deep learning-based recommendation system for healthcare. Performance metrics are MAE & RMSE	Inefficient privacy results	Need to improve results by resolving security features
Archenna et al., [11] (2017)	Big data-based Healthcare recommendation system The performance metric is accuracy	Security and reliability i ssues	Need to add security features
Sharma et al., [15] (2017)	Information retrieval approach for healthcare recommendation system. The performance metric is accuracy	Limited features of diseases	Need to add more features of diseases for more efficient results
Shah et al., [16] (2020)	Healthcare system based on the deep learning framework. Performance metrics are Recall, Precision, Accuracy & F-score	Less dataset size	Need to use more datasets for efficient results

Continued Table 1

Author References	Proposed Work	Outcomes & Limitations	Scope for future work
Sornalakshmi et al., [17] (2020)	Healthcare system based on Apriori algorithm. Performance metrics are Accuracy. Execution time	Reliability issues	Need to use different optimization algorithms for efficient results
Gebremeskel et al., [19] (2019)	Dynamic data handling approach. Performance metrics are MAE & RMSE	Anomalies in patient information	Need to overcome anomalies in the patient dataset
Yoo et al., [20] (2019)	Recommendation system based on mining. Performance metrics are Entropy & Gain	Classification issues	Implement enhance feature classification model
Deng et al., [21] (2019)	Collaborative filtering-based Healthcare system. The performance metric is recall	Challenging implementation	Stage analysis on classified data.
Hui Yuan et al., [26] (2018)	A health recommendation system based on hybrid technique. The performance metric is Precision	Offline collected information utilised only	Detect only a few diseases Efficient for small datasets only
Gujar et al., [27] (2018)	Machine learning-based health recommender system. The performance metric is accuracy	Limited dataset	To predict and evaluate the health of real-time cases

PROPOSED METHODOLOGY

In this study, the proposed methodology presents a system consisting of three different levels pertaining to healthcare entities. These levels, namely data collection, data execution, and output, aim to facilitate improved decision-making for doctors and patients. The primary objective of the proposed system is to assist patients in making timely clinical decisions regarding their medical treatment. The various stages involved in the suggested methodology are depicted in Fig. 1. The first step involves pre-processing the patient data to handle outliers and address missing data. Subsequently, the feature selection method is applied to the

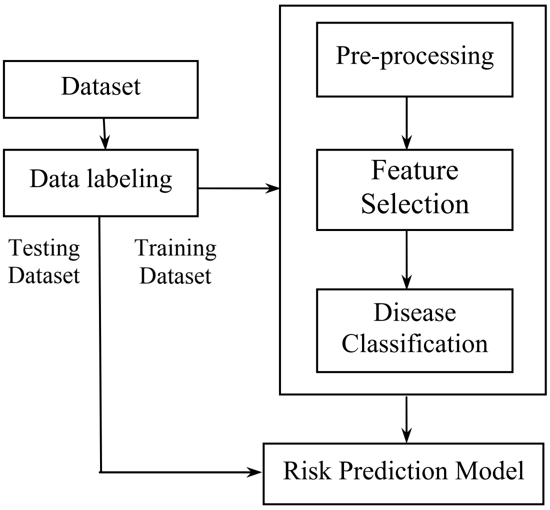


Fig. 1. Architecture for heart disease prediction model

datasets during the classification stage, and the appropriate feature sets are chosen. These selected features are then utilized to predict one of the four predetermined classes of heart disease. The predicted information is subsequently combined with the patient's medical record to offer general medical advice based on the risk level of the disease.

Data Processing

In this phase, data collected data from different recourses undergoes a preprocessing phase. Soft computing techniques have become an essential part of meaningful analysis and obtaining optimal results. To improve the training model's performance, we eliminated unnecessary data like missing values, repeating records and outliers. At the top layer, the collected data is initially cleaned and processed, to enhance the presentation and quality of the data used for model development. To identify outliers, missing values, and irrelevant data, we have employed the numerical cleaner filter technique [28]. The processed data was utilized in the feature selection process. Once the data labeling process is completed, the data is subsequently partitioned into two segments: the training dataset and the testing dataset. The training set is utilized for training the classification model, while the testing set is employed to assess the model's performance.

Feature Selection

In the feature selection phase, a subset of highly distinguishing features must be selected for the diagnosis of diseases. In this process discriminating features are selected for different available classes [30]. The datasets used in this study comprise 20 features, among which only a few are pertinent for decision-making in the disease classification process. To reduce the feature vector to a more manageable sample size, a feature selection technique is employed, consisting of two phases.

In the first phase, an attribute selection technique is utilized to evaluate the features present in the datasets. This technique helps assess the relevance and importance of each feature.

In the second phase, we have employed a search technique to select the optimal set of classification models by systematically exploring different combinations of features. The goal is to identify the most effective combination of features that yields the best performance for the classification task. The proposed selection method in this study is information gain-based, wherein the entropy for each class is evaluated [30]. Features that are selected with higher entropy values are more informative and have a greater contribution to the decision-making process. For taking the computing decision selected highest Info Gain is calculated as followed

$$\text{Info Gain}(X) = \text{Info}(Y) - \text{Info}_{\text{A}}(\text{Y}),$$

X = Investigation feature; Y = Featured Dataset.

In dataset D to classify a feature Info required which is calculated as

$$\text{Info }(I) = -\sum_{i=1}^{n} P_i \log_2(P_i)\,,$$

where n = total number of classes; P = probability; D = dataset.

The selection of features based on their information gain is accomplished through an information gain-based technique in this study. This technique is applied using class labels as a basis, followed by a ranker search method. The purpose of this technique is to determine the relevance of features in the classification task and assign them a ranking accordingly.

Prediction

In the next step selected features are classified for the prediction of diseases after being mapped onto the training model. The classification of diseases is structured as a multi-class issue, where patient data is classified into four primary categories, each representing a distinct disease type.

In order to facilitate the training process, we have employed a classification algorithm based on multilevel Support Vector Machine (SVM) in this study. This algorithm functions by utilizing an accurate function and high dimension to separate class data using a hyper-plane. SVM is optimized for multiple classes to deal with real-world issues. In multilevel classification, pair-wise classification of SVM is used to train the given set of data for each pair of given classes.

Once the model is trained, to evaluate the performance and efficiency of the proposed approach we have applied the trained model to the testing dataset. The prediction model results are assessed using three primary metrics, namely accuracy, sensitivity, specificity and ROC, to determine its performance.

Risk prediction and recommendation model for heart disease

The main goal of this study is to develop a recommendation system that can provide accurate recommendations based on the severity of diseases related to heart. The proposed recommendation model evaluates the patient's data to predict the level of risk and determine the severity of the disease. The algorithm for the proposed model is given below:

Input: P = Prediction of Disease, D = Dataset of patients'

Output: Recommendation R: (1, 2, 3, and 4)

(1: No recommendation, 2: Need to normal exercise, 3: Need to visit doctor, 4: Need to get hospitalized and have proper treatment)

```
$X = (x_1, x_2, x_3)$,  {$X$ represent critical feature set}
($x_1$: Cholesterol, $x_2$: Blood Pressure, $x_3$: Blood sugar)
$Y$ = (Critical, Medium, Normal) {Let $Y$ represent severity range of $X$}
Let W represent the weight of $X$
 Kb = $X$, $W$, $Y$       (Kb= Knowledge base)
for each disease $P$ and info from $k$, $d$
Calculate Probability, Prob ($P$) and Prob ($P$)
Prob ($P$): occurrence of disease, Prob ($P$): absence of disease
R = Prob($P$)/Prob($P$)  {$R$=Estimate Risk}
If $Y$ == Critical then,
For $R$< or > 1 AND for Prob <or > 0.30
Indentify $S$ and compute final score;  {$S$=Score}Final Score = $\sum_{i=1}^{m} S_i(W_i)$
End
Else if $Y$ = medium then,
For $R$> or <1 AND for prob < or > 0.30 then,
$FS$ (Final Score) = $\sum_{i=1}^{m} S_i(W_i)$
```

```
17. End
18. $FS$ (Final Score) = (0 – 5),
19. 0<$FS$<1.9: $R$ = 1
    2< $FS$ <2.9< $R$ = 2
    2.9<$FS$<3.9< $R$ =3
    4<$FS$<5<$R$ = 4
20. End
21. Return $R$
```

Recommendation

Scopes of parameters are identified after the prediction of diseases which also depends upon ranges and risk factor values of severity. The four major general types of recommendations assigned to patients are:

1. No recommendation.
2. Normal Exercise.
3. Visit to doctor.
4. Need to get hospitalized.

The range of parameters in the feature set for heart disease is x_1 = Blood Pressure x_2: Cholesterol, x_3: Blood sugar.

The recommendation classes with score ranges and parameters ranges are given in Table 2 and Table 3.

Table 2. Class labels and ranges recommendation

Class	Labels	Scores
Class 1	No recommendation	0–0.25
Class 2	Normal Exercise	0.25–0.50
Class 3	Visit to doctor	0.50–0.75
Class 4	Need to get hospitalized	0.75–1

Table 3. Ranges of parameters to check heart disease

Parameters	Weightage	Ranges	
		Critical	Normal
Blood pressure	0.75	>160	120–160
Cholesterol	0.50	>300	200–300
Blood Sugar	0.25	>125	100–124

RESULT AND ANALYSIS

In the following subsection, we provide details about the dataset used and present the outcomes of the proposed system obtained through experimental evaluation.

Dataset: For our study, we incorporated a heart disease-related dataset. Our developed system was trained and tested on a heart disease dataset that is openly accessible in the UCI library at http://archive.ics.uci.edu/ml/datasets/heart+disease. This dataset consists of approximately 1000 patients' health records, with health features described in Table 4.

Table 4. Parameter for the health status of patients

Blood Pressure	
Range of BP (mm Hg)	Risk of disease
90/60 (low)	High
120/80 (Normal)	Fit (No Disease)
140/190 (High)	Very High
Cholesterol	
Range of Cholesterol (mg/dL)	Risk of heart disease
100 to 129 mg/dL	Fit (No Disease)
130 to 159 mg/dL	Border Line
160 to 189 mg/dL	High
190 mg/dL and above	Very High

Output for classification phase

Fig. 2 illustrates the ROC curve representing the performance of the predicted model. The prediction of heart disease achieved an AUC of 0.93 for the MSVM algorithm. The ROC curve of the Random Forest (RF) model is closer to 1, indicating higher accuracy. The curve showcases the pair of specificity and sensitivity values for a specific threshold decision at each point.

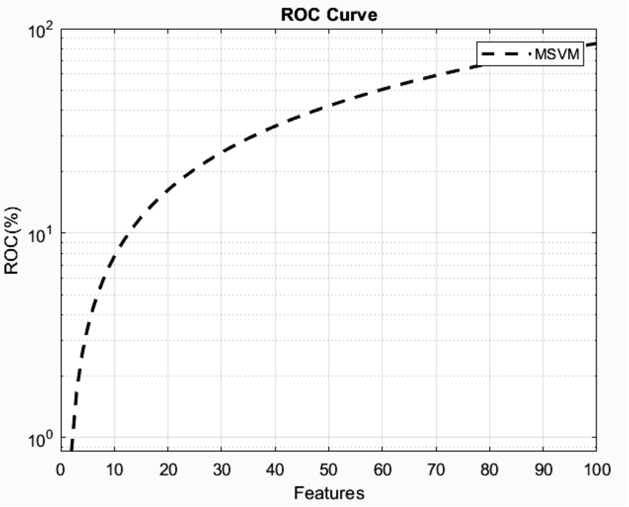


Fig. 2. ROC curve for the prediction algorithm

The performance of the MSVM classifier was evaluated using various performance metrics, primarily accuracy. It was compared to the accuracy of other existing models, namely KNN (85.1%), Naïve Bayes (89.7%), and neural network (91.8%). KNN relied on a single parameter, K (number of neighbours), while Naive Bayes utilized two hyperparameters, α and β, for features classification. The neural network employed the sigmoid activation function to optimize parameters [33].

Fig. 3 presents a curve that represents the accuracy of the multilevel Support Vector Machine (MSVM) classification model. The model achieved an accuracy rate of 94.09%, indicating an optimal solution for improving accuracy compared to existing models such as KNN, Naïve Bayes, and neural network, as shown in Figs. 4, 5 and 6 display the specificity and sensitivity rates, respectively. These figures provide insights into the model's performance in terms of correctly identifying true negatives (specificity) and true positives (sensitivity).

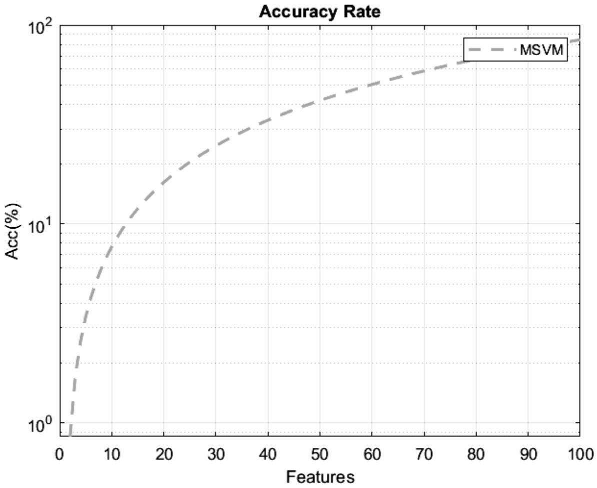


Fig. 3. Accuracy curve for the prediction algorithm

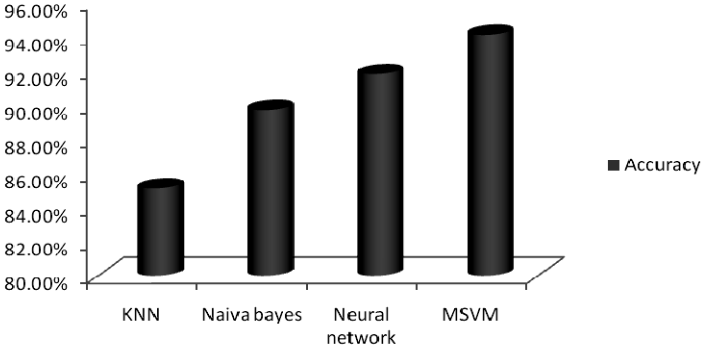


Fig. 4. The Accuracy of different algorithms

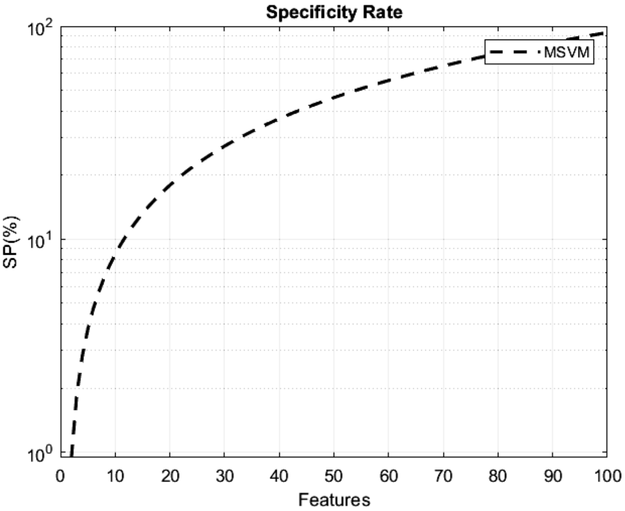


Fig. 5. Specificity Rate

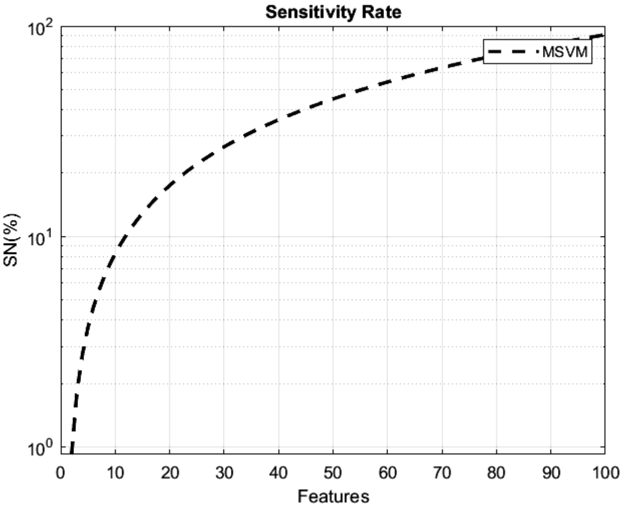


Fig. 6. Sensitivity Rate

The output of the prediction and recommendation phase

In this section, outcome from the prediction and recommendation phase is presented. This research work aims to predict health-related issues in patients and provide recommendations based on the risk and probability of occurrence of diseases. Initially, knowledge is gathered from medical experts and clinical records, and then the weight of each parameter is applied based on its significance. Based on the criticality of parameters different ranges are assigned. Parameters falling in the normal category are ignored.

The following equation is used to evaluate the risk based on the knowledge base database:

$$\text{Risk } (R) = P\,(e)/P\,(e);$$

$$P\,(e) = \text{Probability of disease with abnormality;}$$

$$P\,(e) = \text{Probability of disease without abnormality.}$$

The prediction of the diagnosed disease likelihood is based on the input from the fuzzy system and the classification process and determines the corresponding risk level. By analyzing the given dataset using this method, the resulting values indicate whether the patient's health is at risk or not. The output obtained from the fuzzy model can be found in Table 5.

Table 5. Sample table of outputs of recommendation model

Patient id	Disease	Level	Recommendation
51	Normal	0	No recommendation
57	Heart disease	0.25	Normal Exercise
63	Normal	0	No recommendation
75	Critical Heart disease	0.5	Visit to doctor
81	Highly Emergency	1	Need to get hospitalized

CONCLUSION & FUTURE WORK

The study proposes a framework for predicting and recommending treatments for heart-related diseases using a multilevel decision-making approach. The proposed multilevel classification model, based on the support vector machine, achieves an accuracy rate of 94.09% and provides crucial health recommendations for the early prevention of critical diseases to patients upon disease detection. Implementing these recommendations can help reduce the risk of heart disease and ensure better health outcomes for patients.

In the proposed work, a dataset of around 1000 patients was used. However, it should be noted that the proposed model has limitations, as it relies on clinical validations for any health recommendation decisions.

The proposed model contributes to the medical field by enabling better decision-making for patient healthcare. The analysis of the results focused on accuracy, and we compared the accuracy of our proposed model with three commonly used algorithms (KNN, Naïve Bayes, and Neural Network). We found that the MSVM classification algorithm provided the optimal solution with better accuracy.

In future research, the suggested approach will be evaluated using a real-time dataset. Additionally, the scheme can be expanded by examining the influence of additional characteristics on the recommendation of heart disease, thereby enhancing safety during the validation stage. To achieve this, the implementation of deep learning based technology is necessary. This technology will enable visualization and recording of the learning process, ensuring transparency in the use of deep learning-based techniques.

REFERENCE

1. Y. Li, and T. Beaubouef, "Data Mining: Concepts, Background, and Methods of Integrating Uncertainty in Data Mining," *CCSC: SC Student E-Journal*, 3, pp. 2–7, 2010.
2. *Data Mining Process: Models, Process Steps & Challenges Involved (2021)*. Retrieved 7 July 2021. Available: https://www.softwaretestinghelp.com/data-mining-process/#Steps_In_The_Data_Mining_Process
3. I. Badash et al., "Redefining Health: The Evolution of Health Ideas from Antiquity to the Era of Value-Based Care," *Cureus*, 9(2), e1018, 2017. doi: 10.7759/cureus.1018
4. Robin De Croon, Leen Van Houdt, Nyi Nyi Htun, Gregor Štiglic, Vero Vanden Abeele, and Katrien Verbert, "Health Recommender Systems: Systematic Review," *Journal of Medical Internet Research*, vol. 23, no. 6, e18035, 2021. doi: 10.2196/18035.
5. Q. An, S. Rahman, J. Zhou, and J.J. Kang, "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges," *Sensors*, 23, 4178, 2023. Available: https://doi.org/10.3390/s23094178
6. Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A.A. Yarifard, "Computer aided decision making for heart disease detection using a hybrid neural network-Genetic algorithm," *Comput. Methods and Programs in Biomed.*, 141, pp. 19–26, 2017.
7. H. Yang and J.M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," *J. Biomed. Inform.*, 58, pp. 171–182, 2015.
8. Jhonny Pincay et al., "Health Recommender Systems: A State-of-the-Art Review," *2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG 2019)*, pp. 47–55.
9. K. Subiksha, "Improvement in analyzing healthcare systems using deep learning architecture," in *2018 4th International Conference on Computing Communication and Automation (ICCCA), IEEE, 2018*, pp. 1–4.
10. A.K. Sahoo, S. Mallik, C. Pradhan, B.S.P. Mishra, R.K. Barik, and H. Das, "Intelligence-based health recommendation system using big data analytics," *Big data analytics for intelligent healthcare management*, Academic Press, USA, 2019, pp. 227–246.

11. J. Archenaa, E.A. Mary Anita, “A health recommender system using big data analytics,” *Journal of Management Science and Business Intelligence*, 2(2), pp. 17–24, 2017.
12. K. Vanisree and J. Singaraju, “Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks,” *Int. J. Comput. Appl.*, 19, pp. 6–12, 2015.
13. A. Abugabah, A. AlZubi, F. Al-Obeidat, A. Alarifi, and A. Alwadain, “Data mining techniques for analyzing healthcare conditions of urban space-person lung using meta-heuristic optimized neural networks,” *Cluster Computing*, vol. 23, no. 3, pp. 1781–1794, 2020. doi: 10.1007/s10586-020-03127-w.
14. V. Mudaliar, P. Savaridaasan, and S. Garg, “Disease prediction and drug recommendation android application using data mining (virtual doctor),” *International Journal of Recent Technology and Engineering*, vol. 8, 2019.
15. D. Sharma, G. Singh Aujla, and R. Bajaj, “Deep neurofuzzy approach for risk and severity prediction using recommendation systems in connected Healthcare,” *Transactions on Emerging Telecommunications Technologies*, e4159, 27 October 2020.
16. A. Shah, X. Yan, S. Shah, and G. Mamirkulova, “Mining patient opinion to evaluate the service quality in healthcare: a deep-learning approach,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 7, pp. 2925–2942, 2019. doi: 10.1007/s12652-019-01434-8.
17. M. Sornalakshmi et al., “Hybrid method for mining rules based on enhanced Apriori algorithm with sequential minimal optimization in the healthcare industry,” *Neural Computing and Applications*, 2020. doi: 10.1007/s00521-020-04862-2.
18. O.W. Samuel, G.M. Asogbon, A.K. Sangaiah, P. Fang, and G. Li, “An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction,” *Expert Syst. Appl.*, 68, pp. 163–172 2017.
19. G.B. Gebremeskel, B. Hailu, and B. Biazen, “Architecture and optimization of data mining modeling for visualization of knowledge extraction: patient safety care,” *Journal of King Saud University-Computer and Information Sciences*, 2019.
20. H. Yoo and K. Chung, “Mining-based life care recommendation using peer-to-peer dataset and adaptive decision feedback,” *Peer-to-Peer Networking and Applications*, vol. 11, no. 6, pp. 1309–1320, 2017. doi: 10.1007/s12083-017-0620-2.
21. X. Deng and F. Huangfu, “Collaborative Variational Deep Learning for Healthcare Recommendation,” *IEEE Access*, vol. 7, pp. 55679–55688, 2019. doi: 10.1109/access.2019.2913468.
22. A. Sahoo, C. Pradhan, R. Barik, and H. Dubey, “DeepReco: Deep Learning Based Health Recommender System Using Collaborative Filtering,” *Computation*, vol. 7, no. 2, p. 25, 2019. doi: 10.3390/computation7020025.
23. S. Rathore, M. Habes, M.A. Iftikhar, A. Shacklett, and C. Davatzikos, “A review of neuroimaging-based classification studies and associated feature extraction methods for Alzheimer’s disease and its prodromal stages,” *Neuroimage*, 155, pp. 530–548, 2017. doi: 10.1016/j.neuroimage.2017.03.057.
24. Carolyn Petersen et al., “Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper,” *J. Am. Medical Informatics Assoc.*, 28(4), pp. 677–684, 2021.
25. Khushboo Thaker, “KA-Recsys: Patient-Focused Knowledge Appropriate Health Recommender System,” *SIGIR 2022: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, USA: Association for Computing Machinery, 2022.
26. Hui Yuan and Weiwei Deng, “Doctor recommendation on healthcare consultation platforms: an integrated framework of knowledge graph and deep learning,” *Internet Res.*, 32(2), pp. 454–476, 2022.
27. D. Gujar, R. Biyani, T. Bramhane, S. Bhosale, and T.P. Vaidya, “Disease prediction and doctor recommendation system,” *International Research Journal of Engineering and Technology (IRJET)*, 5, pp. 3207–3209, 2018.

28. S.B. Patil and Y. Kumaraswamy, "Intelligent and effective heart attack prediction system using data mining and artificial neural network," *Eur. J. Sci. Res.*, 31, pp. 642–656, 2009.
29. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian Witten, "The WEKA data mining software: An update," *SIGKDD Explor. Newsl.*, 11, pp. 10–18, 2008.
30. A. Mustaqeem, S.M. Anwar, A.R. Khan, and M. Majid, "A statistical analysis based recommender model for heart disease patients," *Int. J. Med. Inform.*, 108, pp. 134–145, 2017. doi: 10.1016/j.ijmedinf.2017.10.008.
31. Pallavi Chavan, Brian Thoms, and Jason T. Isaacs, "A Recommender System for Healthy Food Choices: Building a Hybrid Model for Recipe Recommendations using Big Data Sets," *HICSS*, pp. 1–10, 2021.
32. Megha Rathi and Vikas Pareek, "Mobile Based Healthcare Tool an Integrated Disease Prediction & Recommendation System," *International Journal of Knowledge and Systems Science*, 10(1), pp. 38–62, 2019. doi: 10.4018/IJKSS.2019010103.
33. N. Priyanka and Pushpa RaviKumar, "Usage of data mining techniques in predicting the heart diseases — Naive Bayes & decision tree," *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. doi: 10.1109/ICCPCT.2017.8074215.

Received 24.04.2023

INFORMATION ON THE ARTICLE

Vedna Sharma, ORCID: 0000-0003-3001-0931, Graphic Era (Deemed to be University) Dehradun, India, e-mail: sharma.vedna@gmail.com

Surender Singh Samant, ORCID: 0000-0001-8619-3779, Graphic Era (Deemed to be University) Dehradun, India, e-mail: surender.samant@gmail.com

БАГАТОРІВНЕВА СИСТЕМА ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ПРОГНОЗУВАННЯ ТА РЕКОМЕНДАЦІЙ ЩОДО ЗАХВОРЮВАНЬ, ПОВ'ЯЗАНИХ ІЗ СЕРЦЕМ / Ведна Шарма, Сурендер Сінгх Самант

Анотація. Точне прогнозування проблем, пов'язаних зі здоров'ям, є серйозною проблемою в галузі охорони здоров'я, причому серцево-судинні захворювання становлять особливо загрозу в глобальній проблемі охорони здоров'я. Точне прогнозування та рекомендації щодо серцево-судинних захворювань мають вирішальне значення для надання своєчасного та ефективного лікування. Основною метою цього дослідження є розроблення моделі класифікації, здатної точно ідентифікувати захворювання серця та надати відповідні рекомендації для пацієнтів. Запропоновано систему, яка застосовує багаторівневий механізм класифікації з використанням опорних векторних машин. Він спрямований на класифікацію захворювань серця шляхом аналізу життєво важливих параметрів пацієнта. Ефективність запропонованої моделі оцінено шляхом її тестуванням на наборі даних, який містить записи пацієнтів. Сформовані рекомендації ґрунтуються на всебічному оцінюванні тяжкості клінічних проявів, які демонструють пацієнти, включно з пов'язаним ризиком як клінічних ознак, так і самого захворювання. Прогнози оцінено за трьома показниками: точність, специфічність і крива робочих характеристик приймача. Запропонована модель класифікації Multilevel Support Vector Machine (MSVM) досягла рівня точності 94,09% у виявленні тяжкості серцевих захворювань, що робить її цінним інструментом у галузі медицини для надання своєчасної діагностики та рекомендацій щодо лікування. Запропонована модель дає багатообіцяльний підхід для точного прогнозування захворювань, пов'язаних із серцем, і засвідчує потенціал методів програмного обчислення в сфері охорони здоров'я. Подальші дослідження можна зосередити на удосконаленні підвищення точності та застосовності запропонованої моделі.

Ключові слова: охорона здоров'я, захворювання серця, модель класифікації, методи навчання.

UDC 519.254:330.47
DOI: 10.20535/SRIT.2308-8893.2023.4.02

METHODOLOGY OF THE COUNTRIES' ECONOMIC DEVELOPMENT DATA ANALYSIS

V.V. DONETS, V.Y. STRILETS, M.L. UGRYUMOV, D.O. SHEVCHENKO, S.V. PROKOPOVYCH, L.O. CHAGOVETS

Abstract. The paper examines the issue of improving the methods of identification of economic objects and their analysis using algorithms of intelligent data processing. The use of the developed methodology in the economic analysis allows for improvement in the quality of management. It can be the basis for creating decision support systems to prevent potentially dangerous changes in the economic status of the research object. In this work, an improved method of c-means data clustering with agent-oriented modification is proposed, and a radial-basis neural network and its extension are proposed to determine whether the obtained clusters are relevant and to analyze the informativeness of state variables and obtain a subset of informative variables. The effect of applying data compression using an autoencoder on the accuracy of the methods is also considered. According to the results of testing of the developed methodology, it was proved that the probability of incorrect determination of the state was reduced when identifying the states of economic systems, and a reduced value of the error of the third kind was obtained when classifying the states of objects.

Keywords: machine learning, digital development, fuzzy clustering, radial basis neural networks, logistic regression, analysis of variables informativeness.

INTRODUCTION

Analysis of the state of economic systems requires taking into account a large number of factors that have a stochastic nature of development and high dynamism. Continuous monitoring allows taking into account the influence of these factors and maintaining the stable functioning of economic systems in conditions of constant global fluctuations [1]. Machine learning methods make it possible to evaluate these factors, their possible and real impact on macroeconomic processes. The use of machine learning algorithms provides early consideration of the effects of factors that may threaten the stability of economic systems [1].

The use of intelligent methods for the analysis of collected economic data allows to automate the solution of many problems in the management of economic processes [1], which significantly increases its quality and efficiency. Automated systems of economic analysis are used as decision support systems to prevent potentially dangerous changes in the state of economic systems [1; 2]. Existing information systems of economic analysis have modules for solving problems of clustering, classification or forecasting of received data, based on machine learning methods, which allow to improve the accuracy of received decisions.

The aim of the study is to improve the quality of data stratification in the information analysis of economic systems by developing a methodology that includes methods of clustering, classification and analysis of the informativeness of economic data. The scientific objective of the study is to improve the existing methods of economic data analysis through the introduction of an agent-oriented modification of the clustering method and radial basis neural networks for analyzing the informativeness of state variables. The proposed methods are expected to reduce the probability of erroneous determination of the state in the analysis of the economic system, thus the value of the third-order error in the classification of its state will be reduced.

STATEMENT OF THE RESEARCH PROBLEM

The data obtained as a result of the study of the economic system can be presented in the form:

$$X = \{x_{im}\},$$

where $i = \overline{1, N}$, $m = \overline{1, M}$, X — matrix representing the data sample for analysis; N — number of objects; M — dimension of space.

The problem of data analysis that characterizes the state of the economic system consists of solving a sequence of problems:

– division of a set of data into sets that are similar according to certain characteristics — the task of clustering;

– determination of the current state of the economic system based on a set of characteristics — the task of classification;

– determination of a set of features that best describe the state of the economic system — the task of selecting informative features (reduction of the space of features).

Let's consider the methods of solving each of the problems.

FUZZY DATA CLUSTERING METHOD

For some known set of valid clusters Y it becomes necessary to split the input data X to $|Y|$ subsets (clusters, classes), so that each cluster consists of objects that are close by some metric, or distant by another. Thus, each object will be assigned to the y-th cluster.

The result of the clustering algorithm [3] will be the application of the function $cluster: X \rightarrow Y$, which matches each object in the input set $x \in X$ matching an object from a set of clusters. Usually, plural $y \in Y$ known in advance for a non-hierarchical approach, or determined in the process for a hierarchical approach. Therefore, the question of determining the optimal number of clusters, as one of the parameters determining the final quality of clustering, often arises.

Let's define the distance between cluster objects as a metric for cluster analysis. Then we define the degree of similarity of objects as the reciprocal of the inter-element distance. Among the works devoted to cluster analysis, can be found a large number of possible metrics for determining the inter-element distance or degree of similarity. The most widespread metric is based on the Euclidean distance, which is a special case of the Minkowski distance [4] with the value of the parameter $\varepsilon = 2$. Generalized Minkowski metric:

$$d_\varepsilon(x_i, x_j) = \sqrt[\varepsilon]{\sum_{m=1}^{M} \left| x_{im} - x_{jm} \right|^{\varepsilon}} .$$

The c-means fuzzy clustering method allows fuzzy distribution of objects into clusters or classes. In the c-means method, the object belongs to all clusters, but with a certain value of cluster membership [5].

In the method of fuzzy clustering [6], the membership matrix of elements to a cluster is calculated according to the assumption of a normal distribution of data according to the formula:

$$w_{ij} = \frac{N(d(x_i, c_j) \mid \mu = 0, \sigma_j)}{\sum_{i=1}^{P_j} N(d(x_i, c_j) \mid \mu = 0, \sigma_j)},$$

where x_i — i-th element of the set, $i = (1; P_j)$; c_j — j-th cluster center; $d(x_i, c_j)$ — distance between points x_i and c_j; $N(d(x_i, c_j) \mid \mu = 0, \sigma_j)$ — probability density of a normal distribution at a point $d(x_i, c_j)$.

The cluster centers are adjusted according to the formula

$$c_j \frac{\sum_{i=1}^{P_j} w_{ij} x_i}{\sum_{i=1}^{P_j} w_{ij}} . \tag{1}$$

The center adjustment process continues until the loss function is minimized:

$$loss = \sum_{j=1}^{K} \sum_{i=1}^{P_j} d(x_i, c_j)^2 w_{ij} \to \min , \tag{2}$$

or on the condition of reaching some limitation on the number of iterations, or the required classification quality.

Among the important disadvantages of the c-means method are the inability to divide the space with a complex shape of target clusters that go beyond simple *M*-dimensional spheres, and an insufficient level of robustness to noise [5; 7].

For data from real problems, both a complex distribution of object parameters and a high dimensionality of the input data are inherent, which in turn determines the complex form of *M*-dimensional target clusters. Therefore, for the usual method of fuzzy clustering and many of its modifications, clustering with high accuracy is not possible. A modification of the distance metric (together with the membership metric) is proposed in [8]. An interesting approach is the assumption of the Cauchy distribution and the use of the Mahalanobis distance, which were proposed in [9; 10]. Mahalanobis distance was used to improve the calculation algorithm that prevents degeneracy of the inverse matrix [11]:

$$MD(x_i, c_j) = \sqrt{(x_i - c_j)^T \hat{\Sigma}_j^{-1} (x_i - c_j)} ,$$

where $\hat{\Sigma} = \Sigma + \lambda\Sigma$ — is the regularized covariance matrix; λ — is a constant greater than zero.

Taking into account the assumption of the Cauchy distribution in the data, the expression for calculating the value of belonging to a certain cluster [5] has the form:

$$w_{ij} = \frac{\rho(x_i, c_j)}{\sum_{i=1}^{P_j} \rho(x_i, c_j)}, \quad \rho(x_i, c_j) = \left(\pi\eta \left[1 + \frac{MD^2(x_i, c_j)}{\eta^2} \right] \right)^{-1}. \qquad (3)$$

Solving the clustering problem for clusters of complex M-dimensional form using Gaussian mixture models was considered in works [12; 13], using the derivative in [14] and using the Mahalanobis distance in [5; 9]. According to the obtained results, an improvement in clustering accuracy is noted, but the problem of spatial separation and overuse of input data dependence occurs.

In works [5; 15], the possibility of taking into account the relative entropy of the data distribution was considered when using the c-means method, but the Euclidean distance was chosen as the metric of the distance between the objects of the sample, which reduced the computational load, but did not take into account the entropy of the data.

To overcome the difficulties of using the basic method of fuzzy clustering and its modifications based on Mixture and Gaussian mixture models on data with a complex shape of M-dimensional target clusters [12], which is based on an attempt to take into account the entropy of clusters [15] and the Kullback–Leibler distance [16], it was proposed to improve the clustering method.

The Kullback–Leibler distance is an asymmetric measure of the informational difference between two probability distributions. This measure has proven itself well in methods of information processing in physical systems and statistics [16].

According to the previous definition $x_{im} \in X$ — is the m-th state variable of the i-th vector of the input data sample, where $m \in [1, M]$, M — dimension of the state vector. Let's define $f_s \in F$ as the s-th object function from the vector of object functions $s \in [1, S]$, where S — the dimension of the object functions vector. Then $M_\alpha(f_s)$ and $M_\alpha(x_{im})$ are mathematical expectations of f_s and x_{im} respectively. According to this definition $D(f_s)$ and $D(x_{im})$ — dispersion of the relevant variables, and $\sigma(f_s)$ and $\sigma(x_{im})$ — standard deviation. Variance and standard deviation of conditional dependence of f_s from x_{im} an be determined by formulas:

$$D(f_s | x_{im}) = var(M_\alpha(f_s(x_{in-m}))), \ \forall n, \ n \neq m, \ x_{in} = const; \qquad (4)$$

$$\sigma(f_s | x_{im}) = \sqrt{D(f_s | x_{im})}. \qquad (5)$$

Using expression (4), we obtain estimates of informative state variables:

$$\beta(f_s) = \frac{D(f_s | x_{im})}{E(f_s)},$$

where $E(f_s)$ — signal energy.

From (5), we get the influence coefficient (signal to noise ratio):

$$\varphi_{sm} = SNR(f_s | x_{im}) = \frac{\sigma(f_s | x_{im})}{\sigma(x_{im})}.$$

In [16], the Kullback–Leibler entropy is defined as follows:

$$D_{KL}(f_s, x_i) = \sum_{m=1}^{M} \rho(x_{im} \,|\, f_s) \log_2 \left(\frac{\rho(x_{im} \,|\, f_s)}{\rho(x_{im})} \right).$$

Mutual informative dependence is then determined by the formula:

$$H_{sm} = \frac{1}{2} \log_2 SNR^2(f_s \,|\, x_{im}) = \frac{1}{2} \log_2 \left(\beta(f_s) \frac{E(f_s)}{D(x_{im})} \right).$$

In the proposed method, we replace the loss function (2). Instead, we will get a formula for determining mutual informative dependence, which will be a function of clustering quality assessment, that is, a function of losses in the developed method of fuzzy clustering:

$$H(X,Y) = -\frac{1}{\sum_{j=1}^{k} P_j} \sum_{j=1}^{k} \left[P\left(Y_j^{(t+1)} \times \sum_{i=1}^{P_j} D_{KL}(x_i, Y_j^{(t+1)})` \right) \right] \to \min ,$$

where Y_j — state variables belonging to the *j*-th cluster.

AGENT-ORIENTED MODIFICATION OF THE CLUSTERIZATION METHOD

To overcome the non-priority problem, an agent-oriented modification was developed for the classical method of fuzzy clustering considering the *M*-dimensional spatial shape [3; 5], which is considered below.

Let`s introduce special notations for the developed method of fuzzy clustering: *X* — agents, elements of the input sample, *C* — centers of clusters, then X_i — agents, cluster elements, *Z* — agents clusters. According to the agent-oriented approach, the elements-vectors of the input sample and the clusters are agents, these agent-elements choose the cluster agents closest to them, which they join according to a pre-specified metric, thus forming cluster agents. The number of cluster agents is determined by minimizing the loss function. According to the previous definition: the input sample partitioned into clusters is $X = \{P_j\}$, where $j \in (1, K)(1,\ K)$, $N = \sum_{j=1}^{K} |P_j|$ — the number of elements in the input sample; P_j — set of elements belonging to the *j*-th cluster; *K* — number of clusters. Than $x_{ij} \in P_j$ — the *i*-th element of the *j*-th cluster.

Four metrics were chosen to compare the possibilities of spatial separation of clusters and computational efficiency:

$$d(x_{ij}, c_j) = \begin{cases} d_1(x_{ij}, c_j), \\ w_{ij}^{-1} d_1(x_{ij}, c_j), \\ -D_{KL}(x_{ij}, c_j), \\ p(x_{ij}, c_j^{t-1}) * \log_2 p(x_{ij}, c_j^{t}), \end{cases} \tag{6}$$

where $d_1(x_{ij}, c_j)$ — Manhattan distance; $w_{ij}^{-1} d_1(x_{ij}, c_j)$ — Mahalanobis distance with the inverse of the membership function; $-D_{KL}(x_{ij}, c_j)$ — Kullback–Leibler divergence; $p(x_{ij}, c_j^{t-1}) * \log_2 p(x_{ij}, c_j^{t})$ — cross entropy.

Having the distance to determine the inter-element distance, we will get an expression for determining the cost function for each cluster, that is, the average measure of the intraclass distance:

$$cl\_loss(P_j) = \frac{1}{|P_j|} \sum_{i=1}^{|P_j|} d(x_{ij}, c_j) . \tag{7}$$

Then, using expression (7), we obtain the general cost function for evaluating the current quality of clustering:

$$loss(X^t) = \frac{1}{K^t} \sum_{j=1}^{K^t} cl\_loss(P_j) . \tag{8}$$

By combining the classical method of fuzzy clustering with the agent-oriented approach described above, we will obtain a statement of the research problem, according to which it is necessary to determine the number of clusters and such a distribution of elements by clusters that the value of the cost function is minimal:

$$\begin{cases} A = [K^t, X^t], \\ \hat{A} = \arg\min (loss(X^t)). \end{cases}$$

According to the classical clustering method, cluster centers are optimized according to expression (1), and the membership matrix for adjustment is calculated according to expression (3) taking into account the Cauchy distribution assumption. We formulate the clustering algorithm, defined according to the agent-oriented approach, as follows:

1. Determine some initial number of cluster agents $K^t > K$, that is more than the target number of clusters, and set a limit on the number of elements in each cluster $|P_j^t| = N / K^t$ and choose randomly K^t centers of clusters $\{c_j\}$.

2. Select one of the inter-element distances (6) $|P_j^t|$ of the closest elements to each cluster, that is, to form cluster agents P_j^t.

3. For each cluster, calculate the value of the parameters $\rho(x_{ij} | P_j^t)$ distribution and the values of the membership matrix according to expressions (3), and according to expression (1) adjust the cluster centers.

4. To each center of the cluster according to the selected measure $d(x_{ij}, c_j)$ to choose $|P_j^t|$ new agents-elements.

5. For each cluster agent, according to expression (7), determine the value of the cost function (or the average inter-element distance) $cl\_loss\left(P_j^t\right)$.

6. To estimate the current quality of clustering by the loss function according to expression (8). In the case of the operation mode of the algorithm in the automatic search for the optimal number of clusters, and the increase in the value of the cost function, stop the algorithm.

7. To select agent-clusters and discard the agent-cluster with the highest value $cl\_loss(P_j^t)$.

8. To determine the new number of clusters $K^{t+1} = K^t - 1$ and the new number of cluster elements $| P_j^{t+1} | = N / K^{t+1}$.

9. Return to stage 2, if $K^t > K$.

CLASSIFICATION METHOD BASED ON MULTIPLE LOGISTIC REGRESSION

To solve the problem of multiclass classification in the case of spatially separated data, it is proposed to use a radial basis neural network (RBFN) with multiple logistic regression. The application of the RBFN model for multiclass classification will allow checking the assumptions about the correctness of the cluster definition and testing the model's ability to generalize.

RBFN structure: H_0 inputs for each of the parameters, H_1 neurons of the first layer and H_2 output neurons. We define the vector of input data for the *k*-th layer of the neural network (or the vector of output data for the *k*-1 layer) as $\vec{Y}^{(k)} = [Y_1^{(k)}, \ldots, Y_{H_1}^{(k)}]^T$, we define the vector of coordinates of the cents of the activation function for the hidden layer as $\vec{c}_j = [c_{j1}, c_{j2}, \ldots, c_{jH_0}]^T$, where $j = 1..H_1$, and the vector specifying the window width of the activation function of the *j*-th neuron of the hidden layer is defined as $\vec{\sigma}_j = [\sigma_{j1}, \sigma_{j2}, \ldots, \sigma_{jH_0}]^T$. Then the activation function for the neurons of the hidden layer will look like this:

$$\varphi_j = (\vec{Y}_p^{(0)}, \vec{c}_j, \vec{\sigma}_j) = exp\left(-\frac{1}{2} \sum_{h=1}^{H_0} w_{ij} Z_{pjh}^2 \right) \equiv \varphi_{pj},$$

where $Z_{pjh} = \dfrac{Y_{ph}^{(0)} - c_{jh}}{\sigma_{jh}}$; w_{ij} — weighted connection between the *i*-th neuron of the output layer and the *j*-th neuron of the input layer.

Multiple logistic regression [17] is used as the activation function of the output layer, the outputs of which are defined as:

$$\vartheta_j = \frac{\exp(\gamma_j)}{\sum_{k=1}^{H_2} \exp(\gamma_k)}, \text{ де } \gamma_\text{j} = \sum_i^{H_1} \varphi_i w_{ij}.$$

A hybrid algorithm was used for training the RBFN, which includes 2 steps, the repetition of which usually leads to fast training of the network, especially if the parameters are successfully generated [18]:

1) selection of linear network parameters (weights) using the pseudo inversion method;

2) optimization of nonlinear parameters of activation functions (window centers and widths).

If there are *P* training pairs $(\vec{Y}_p^{(0)}, \vec{d}_p)$, $p = 1..P$ and fixing the specific values of the centers and window widths of the activation functions, we get a system of equations:

$$\Phi\vec{w}_i = \vec{d}_i,\ i = 1..H_2,$$

where $\Phi = [\varphi_{pj}]$, $p = 1..P$, $j = 0..H_1$, $\varphi_{p0} = 1$, $\vec{w}_i = [w_{i0}, w_{i1}, ..., w_{iH_1}]^T$, $\vec{d}_i = = [d_{0i}, d_{1i}, ..., d_{pi}]^T$.

Vector $\vec{w}_i$ can be determined in one step using pseudo matrix inversion Φ: $\vec{w}_i = \Phi^+\vec{d}_i$, which in practice is calculated using the decomposition of eigenvalues.

At the second stage of the algorithm, when fixing the weights, the excitation signal passes through the network to the initial level, which allows to calculate the error value for the sequence of vectors $\{\vec{Y}_p^{(0)}\}$. After that, there is a return to the hidden layer. The gradient vector of the selection function according to the specific variable cents and window widths is determined by the error value: $\| \vec{Y}^{(2)} - \vec{d} \|_{L_2}$.

Algorithm for forming the "coverage zone" by radial basis functions of k-neighbors $\sigma_{jh}^2 = \Sigma_j = \frac{1}{K}\sum_{k=1}^{K}\sum_{h=1}^{H_0}(c_{jh} - c_{kh})^2$, $k = 1..K$, $K \in [3,5]$ was used to determine the values of the window widths, which helped reduce the training time of the RBFN.

CHARACTERISTICS INFORMATIVENESS ANALYSIS METHOD

Since it is proposed to use the RBFN network to solve the classification problem, this model can also be used to find the minimum possible subset of informative variables. The input data set can be represented as a Taylor series, keeping only the terms of the first infinitesimal order. For the variance of an arbitrarily obtained linear function of several random variables, the estimate is valid:

$$D_{Y_i} = (\operatorname{grad} Y_i)^T \, \Sigma_S \operatorname{grad} Y_i = \sum_{j=1}^{J}\left(\frac{\partial Y_i}{\partial s_j}\right)^2 \sigma_{S_j}^2 + \sum_{j=1}^{J}\sum_{l=1, l\neq j}^{J} r_{jl}\frac{\partial Y_i}{\partial s_j}\frac{\partial Y_i}{\partial s_l}\sigma_{S_j}\sigma_{S_l},$$

where Σ_S — covariance matrix of variables $S_1; S_2$, σ_{S_1} — standard deviation; r_{j1} — correlation coefficient between variables S_1 and S_2.

Then the standard deviation and variance of the RBFN output can be estimated according to the architecture chosen for it, and from them determine the energy of the signals by the expression [18]:

$$E_i = \sum_{h=1}^{H_0}\left| D_{Y_i^{(2)}|Y_h^{(0)}} \right|,$$

where $D_{Y_i^{(2)}|Y_h^{(0)}} = \left(\frac{\partial Y_i^{(2)}}{\partial Y_h^{(0)}}\right)^2 \sigma_{Y_h^{(0)}}^2 + \left(\sum_{n=1, n\neq h}^{H_0} r_{hn}\frac{\partial Y_i^{(2)}}{\partial Y_n^{(0)}}\sigma_{Y_n^{(0)}}\right)\frac{\partial Y_i^{(2)}}{\partial Y_h^{(0)}}\sigma_{Y_h^{(0)}}$.

Then the coefficient of informativeness of the variables (the weight of the contribution of $Y_h^{(0)}$ in to $Y_i^{(2)}$) is defined by the expression:

$$\beta_{ih} = \frac{\left| D_{Y_i^{(2)} | Y_h^{(0)}} \right|}{E_i}.$$

DATA PRE-PROCESSING METHODS

In machine learning problems, it has become common practice to use data pre-processing methods (normalization, cleaning from anomalies, and dimensionality reduction) to improve the quality of problem solving [19]. Three methods of the scikit-learn, Python library were used for data *normalization*:

– *RobustScaler* scales parameters with robustness to statistical outliers.

– *StandardScaler* (*Z*-score normalization). Reduces the mean and scales to unit variance.

– *MinMaxScaler* (min-max normalization). Each parameter is scaled and translated individually by the estimator so that it falls within a given range, for example [0,1].

The detection of unusual elements, events, or observations that are significantly different from the main body of data and do not correspond to a well-defined definition of normal behavior is called the process of anomaly detection [20]. Data cleaning techniques remove values that have been identified as outliers and based on anomaly detection.

Two outlier detection methods from the scikit-learn library were used:

– *Interquartile Range* (IQR). By dividing the data set into quartiles, it is used to measure variability;

– *Isolation forest*. The method uses isolation to find anomalies (how far a data point is from the rest of the data) [21; 22].

The *dimensionality reduction* process aims to provide a lower-dimensional representation of the original data set while preserving its important characteristics. Separate scikit-learn and PyTorch libraries were used for dimensionality reduction. Three methods were used:

– *T-distributed Stochastic Neighbor Embedding* (t-SNE) [23];

– *Principal Component Analysis* (PCA) the method is based on SVD, it reduces the dimensionality of the data well [24].

– *Autoencoder*. Is a certain type of feed-forward neural network where the input matches the output. It compresses the input data into a bottleneck (lower dimensional data) and then reconstructs the output data from that representation. The bottleneck is the target compact summation or dimensionality reduction of the input data, also called the latent space representation.

APPLICATION OF METHODOLOGY FOR COUNTRIES DIGITAL DEVELOPMENT DATA ANALYSIS

The developed methodology was tested to identify the state of digital development of the countries of the world. For the classification (positioning of countries) regarding the level of their digital development, the hypothesis of the existence of homogeneous groups of countries (objects) according to specialized indices was tested. Indices that fully reflect the state of digital development were selected:

– EGIit — Global E-Government Development Index;

– NRIit — network readiness index;

– ICTit — information and communication technologies development index.

By forecasting independent factors — indicators of digital development based on the model, it is possible to estimate the forecast level of social progress of a specific country. The Social Progress Index (SPI) is a combined indicator of the International Research Project The Social Progress Imperative [25; 26] which measures the achievements of the countries of the world in terms of social well-being and social progress. The authors of the study [25; 26] believe that indicators of social development are often considered as an alternative to indicators of economic development. The global e-government development index [26] is an integral indicator that assesses the readiness and capabilities of national government structures in using information and communication technologies (ICT) to provide public services to citizens. The index of network readiness [26] characterizes the level of development of information and communication technologies and the network economy in the countries of the world. Currently, the index is considered one of the most important indicators of the innovative and technological potential of the countries of the world and their development opportunities in the field of high technology and digital economy. The ICT Development Index is a composite index that combines 11 indicators and is used to monitor and compare the development of information and communication technologies (ICT) between countries.

To implement the model, a sample of 115 precedents (observations by country) was collected for 32 variables of the state of social development for each precedent and the 33rd field for the predictive value of the state. The ratio of values of the social progress index SPIt (Social Progress Index) and the average level of income was used to mark the educational sample. All precedents of the sample were distributed according to the respective states:

– “High income” — 45 precedents (I);

– “Upper middle income” — 11 precedents (II);

– “Lower middle income” — 25 precedents (III);

– “Lower income” — 34 precedents (IV).

For this sample, pre-processing of the data was first carried out: normalization and detection of anomalous values. Clustering was performed for the considered economic data, and classification was performed to verify its results. It was decided to use the Kullback–Leibler distance classification method. As a result of its application, an accuracy of 84.3% was achieved, and the value of the flow function was obtained as 0.0117. A matrix of inconsistencies (Table 1) was also constructed to assess the accuracy of the method, as well as graphs of cost function values (Fig. 1) and ROC curves for each of the classes (Fig. 2).

Table 1. The matrix of inconsistencies in the classification of data indicators of the digital development of the countries of the world

Actual class	**Predicted class**			
	I	II	III	IV
I	37	1	0	7
II	1	8	1	1
III	2	0	21	2
IV	0	1	2	31

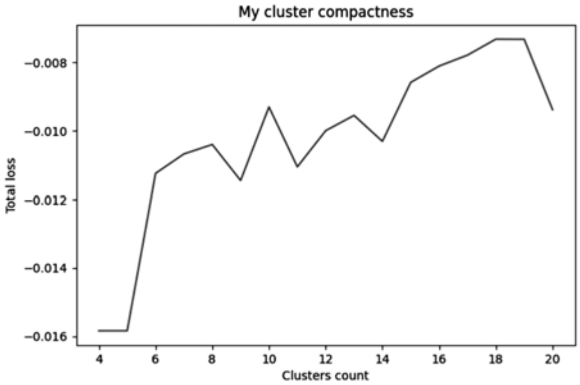


Fig. 1. The ratio of the number of clusters to the value of the cost function for economic indicators of the countries of the world data

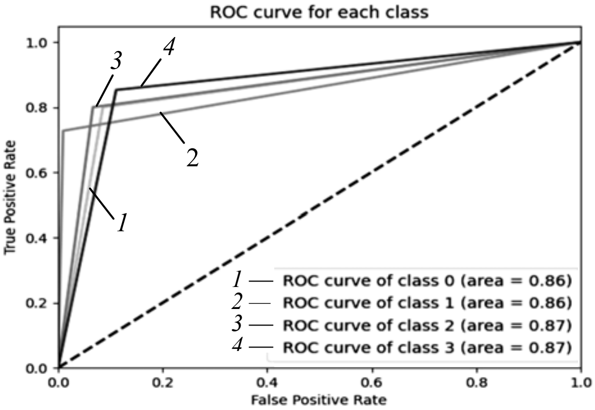


Fig. 2. ROC curves for each of the classes for these economic indicators of the countries of the world

After a series of experiments, it was decided to apply the autoencoder method to reduce the dimensionality of the data with 98% information retention, which made it possible to reduce the dimensionality of 32 to 11 state variables for each case. After this application, an accuracy of 86.9% was achieved, and the value of the cost function became -0.04827. A matrix of inconsistencies (Table 2) was also constructed to assess the accuracy of the method and a graph of the values of the cost function (Fig. 3) and ROC curves for each of the classes (Fig. 4).

Table 2. The matrix of inconsistencies in the classification of compressed data indicators of the digital development of the countries of the world

Actual class	Predicted class			
	I	II	III	IV
I	38	0	0	7
II	0	10	0	1
III	1	0	21	3
IV	0	1	2	31

To carry out *multi-class classification* with the help of RBFN, the data of the digital development of countries with a reduced dimension, processed by the autoencoder method, were used. To test the ability of the model to generalize, the data were divided into test and training samples in the ratio of 20% (22 precedents) and 80% (93 precedents), respectively. Previously, the data sample was normalized.

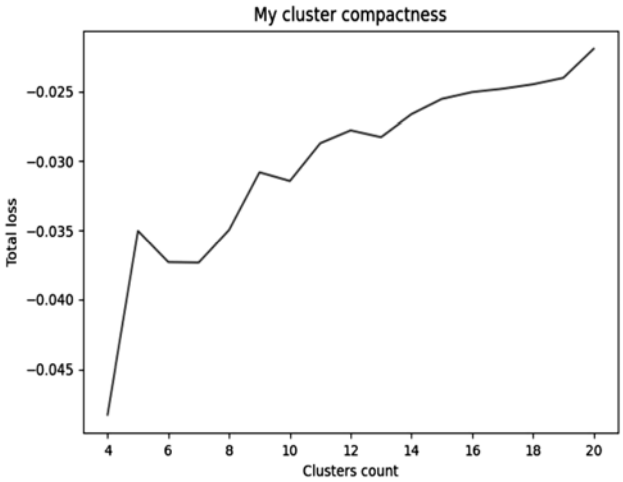


Fig. 3. The ratio of the number of clusters to the value of the cost function for the compressed data of the economic indicators of the countries of the world

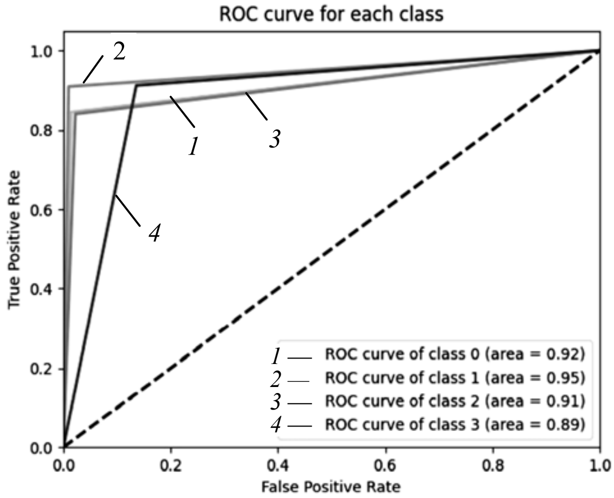


Fig. 4. ROC-curves for each of the classes for compressed data of economic indicators of the countries of the world

RBFN will receive 7 state variables that do not have a defined value at the input, and at the output there will be estimates of state variable values — 4 states. The structure of the proposed RBFN has $H_0 = 7$ inputs for each of the parameters, $H_1 = 90$ neurons of the first layer and $H_2 = 4$ output neurons.

Table 3. Misclassification matrix of the compressed data of the country's digital development indicators of the world

Actual class	**Predicted class**			
	I	II	III	IV
I	8	0	0	1
II	0	2	0	1
III	0	4	1	0
IV	2	0	0	4

As a result of training on the training sample, an accuracy of 83.87%, while on the test sample — 68.18%. To display the test results, a matrix of inconsistencies was constructed for the training sample (Table 3) and a ROC curve was shown (Fig. 5), which has a smaller coverage area (i.e., worse classification ability), because part of the data was used for training, which reduced the ability of RBFN to generalization.

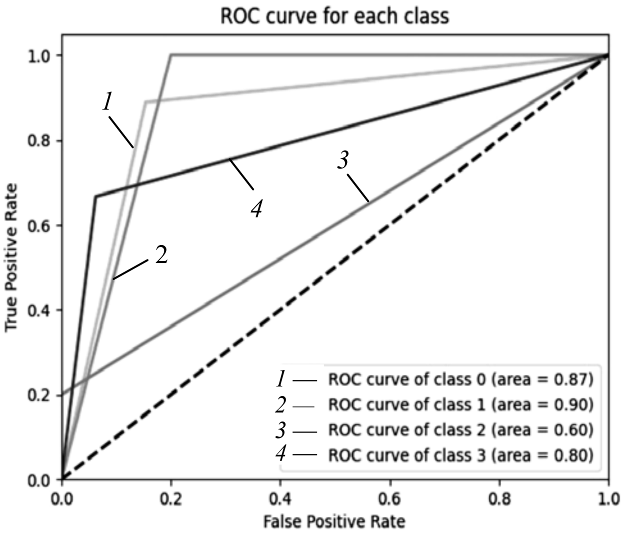


Fig. 5. ROC curves for each of the classes for the PCA test sample of compressed data of indicators of digital development of the countries of the world

An analysis of the sensitivity of the target function was also carried out, i.e. the most informative indicators were determined. The results are shown in Table 4. Based on the results, it can be concluded that a different set of variables is informative for each cluster.

Table 4. Sensitivity analysis of the variable clusters objective functions

Cluster	**Number of precedents**	**Sensitive cluster variables**	**Mathematical expectation of the objective function**
0	45	TII, ICT, HCI	85.33
1	11	TII, ICT, EGI	52.87
2	25	EPI, HCI, OSI	63.47
3	34	HCI, EPI, EGI	73.60

All numerical studies were carried out using the computer program "Nonlinear estimation methods in the multicriterion problems of system's robust optimal designing and diagnosing under parametric apriority uncertainty (methodology, methods and computer decision support and making system" (ROD&IDS), developed by the authors [27].

CONCLUSIONS

The methods of intelligent data flow processing are widely used during the identification of the states of economic objects. The use of new methods will make it possible to supplement the package of available tools for solving current problems with data processing and will make it possible to increase the stability of the methods to the nature of the data and improve the situation with the use of computing resources.

Presented study examines the problem of improving the methods of classification and clustering of countries according to the state of social and digital development. A multiclass classification method based on radial basis neural networks and a data clustering method based on an agent-oriented modification of the c-means method are proposed.

The proposed RBFN uses multiple logistic regression as the last layer for multiclass classification and the training results of an agent-oriented clustering model as input parameters. The peculiarity of the modification of the c-means method is the introduction of elite selection of clusters.

According to the results of the research, the proposed methodology is proposed to be used for the analysis of economic systems to improve the quality of decision-making, but it should be noted that the method requires a qualitatively prepared sample that covers the largest possible space of input parameters for the target classes.

REFERENCES

1. Mei Yang, Ming K. Lim, Yingchi Qu, Du Ni, and Zhi Xiao, "Supply chain risk management with machine learning technology: A literature review and future research directions," *Computers & Industrial Engineering*, vol. 175, January 2023, 108859. Available: https://doi.org/10.1016/j.cie.2022.108859
2. Benjamin Decardi-Nelson and Jinfeng Liu, "Robust Economic Model Predictive Control with Zone Control," *IFAC-PapersOnLine*, vol. 54, issue 3, pp. 237–242, 2021. Available: https://doi.org/10.1016/j.ifacol.2021.08.248
3. M. Schlesinger and V. Hlavac, *Ten lectures on statistical and structural pattern recognition*. Springer, Dordrecht, 2002. doi: 10.1007/978-94-017-3217-8.
4. *Data clustering: algorithms and applications*, Charu C. Aggarwal and Chandan, K. Reddy (ed.). CRC Press, Taylor & Francis Group, 2014.
5. N. Bakumenko, V. Strilets, and M. Ugryumov, "Application of the C-Means Fuzzy Clustering Method for the Patient's State Recognition Problems in the Medicine Monitoring Systems," *CEUR Workshop Proceedings of 3rd International Conference on Computational Linguistics and Intelligent Systems, COLINS 2019*, vol. I, pp. 218–227, 2019, Available: https://www.researchgate.net/publication/338819685
6. R. Winkler, F. Klawonn, and R. Kruse, "Problems of Fuzzy c-Means Clustering and Similar Algorithms with High Dimensional Data Sets," *Challenges at the Interface of Data Analysis, Computer Science and Optimization*, pp. 79–87, 2012. doi: 10.1007/978-3-642-24466-7_9.
7. Christopher D. Prabhakar Raghavan and Hinrich Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.
8. S. Askari, "Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," *Expert Systems with Applications*, vol. 165, article no. 113856, 2020. doi: 10.1016/j.eswa.2020.113856.
9. Xuemei Zhao, Yu Li, and Quanhua Zhao, "Mahalanobis distance based on fuzzy clustering algorithm for image segmentation," *Digital Signal Processing*, vol. 43, pp. 8–16, Aug 2015. Available: https://doi.org/10.1016/j.dsp.2015.04.009
10. Zarinbala M. Zarandia, M.H. Fazel, and I.B. Turksen, "Relative entropy fuzzy *c*-means clustering," *Information Sciences*, vol. 260, pp. 74–97, 2014. doi: 10.1016/j.ins.2013.11.004.
11. V. Strilets, V. Donets, M. Ugryumov, R. Zelenskyi, and T. Goncharova, "Agent-Oriented data clustering for medical monitoring," *Radioelectronic and Computer Systems*, no. 1, pp. 103–114, 2022. Available: https://doi.org/10.32620/reks.2022.1.08
12. Meng Xing, Yanbo Zhang, Hongmei Yu, Zhenhuan Yang, and Xueling Li, "Predict DLBCL patients' recurrence within two years with Gaussian mixture model cluster oversampling and multi-kernel learning," *Computer Methods Programs in Biomedicine*, vol. 226, 107103, 2022. Available: https://doi.org/10.1016/j.cmpb.2022.107103
13. Lynne A. Kvapil, Mark W. Kimpel, Rasitha R. Jayasekare, and Kim Shelton, "Using Gaussian mixture model clustering to explore morphology and standardized production of ceramic vessels: A case study of pottery from Late Bronze Age Greece,"

Journal of Archaeological Science: Reports, vol. 45, 103543, 2022. Available: https://doi.org/10.1016/j.jasrep.2022.103543

14. Meng Yinfeng, Jiye Liang, Fuyuan Cao and Yijun He, "A new distance with derivative information for functional k-means clustering algorithm," *Information Sciences*, vol. 463–464, pp. 166–185, 2018. Available: https://doi.org/10.1016/ j.ins.2018.06.035
15. Xinmin Tao, Ruotong Wang, Rui Chang, and Chenxi Li, "Density-sensitive fuzzy kernel maximum entropy clustering algorithm," *Knowledge-Based Systems*, vol. 166, pp. 42–57, 2019. Available: https://doi.org/10.1016/j.knosys.2018.12.007.
16. K. Møllersen, S. Dhar and F. Godtliebsen, "On Data-Independent Properties for Density-Based Dissimilarity Measures in Hybrid Clustering," *Applied Mathematics*, vol. 7, no. 15, pp. 1674–1706, 2016. doi: 10.4236/am.2016.715143.
17. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Softmax Units for Multinoulli Output Distributions. Deep Learning*. MIT Press, 2016.
18. V.E. Strilets et al., *Methods of machine learning in the problems of system analysis and decision making: monograph*. Karazin Kharkiv National University, 2020, 195 p.
19. Farbod Farhangi, "Investigating the role of data preprocessing, hyperparameters tuning, and type of machine learning algorithm in the improvement of drowsy EEG signal modeling," *Intelligent Systems with Applications*, vol. 15, 200100, September 2022. Available: https://doi.org/10.1016/j.iswa.2022.200100
20. Arthur Zimek and Peter Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(6), 2018. doi: 10.1002/widm.1280.
21. Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, "Isolation-Based Anomaly Detection," *ACM Transactions on Knowledge Discovery from Data*, 6(1), pp. 1–39, 2012. doi:10.1145/2133360.2133363.
22. O.Yu. Lykhach, M.L. Ugryumov, D.O. Shevchenko, and S.I. Shmatkov, "Methods of detecting emissions in test samples during process control in state-based systems," *Bulletin of Karazin Kharkiv National University, ser. "Mathematical modeling. Information Technology. Automated control systems"*, no. 53. pp. 21–40, 2022.
23. L.J.P van der Maaten and G.E. Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research*, 9, pp. 2579–2605, 2008.
24. Ian T. Jolliffe and Jorge Cadima, "Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A," *Mathematical, Physical and Engineering Sciences,* 374(2065), 20150202, 2016. doi: 10.1098/rsta.2015.0202.
25. L. Chagovets, N. Chernova, T. Klebanova, O. Dorokhov, and A. Didenko, "Selective Adaptive Model for Forecasting of Regional Development Unevenness Indexes," *Proceedings of the Workshop on the XII International Scientific Practical Conference Modern problems of social and economic systems modelling (MPSESM-W 2020) Kharkiv, Ukraine, June 25, 2020,* pp. 58–76.
26. L.O. Chagovets, S.V. Prokopovych, S.M. Vozniuk, and V.V. Chahovets, "Conceptual basis of modeling telecommunication development of regions by methods of system analysis," *Municipal economy of cities*, vol. 1, no. 161, pp. 230–240, 2021.
27. Computer program "Nonlinear estimation methods in the multicriterion problems of system's robust optimal designing and diagnosing under parametric apriority uncertainty (methodology, methods and computer decision support and making system)" ("ROD&IDS"): Copyright registration certificate no. 82875 / M.L. Ugryumov, Y.S. Meniaylov, S.V. Chernysh, K.M. Ugryumova (Ukraine). Copyright and related rights. Official bulletin. Ministry of Economic Development and Trade of Ukraine. 2018, no. 51, p. 403.

Received 30.06.2023

INFORMATION ON THE ARTICLE

Volodymyr V. Donets, ORCID: 0000-0002-5963-9998, V.N. Karazin Kharkiv National University, Ukraine, e-mail: v.donets@karazin.ua

Viktoriia Y. Strilets, ORCID: 0000-0002-2475-1496, V.N. Karazin Kharkiv National University, Ukraine, e-mail: viktoria.strilets@karazin.ua

Mykhaylo L. Ugryumov, ORCID: 0000-0003-0902-2735, V.N. Karazin Kharkiv National University, Ukraine, e-mail: m.ugryumov@karazin.ua

Dmytro O. Shevchenko, ORCID: 0000-0002-7897-250X, V.N. Karazin Kharkiv National University, Ukraine, e-mail: dimyich24@gmail.com

Svitlana V. Prokopovych, ORCID: 0000-0002-6333-2139, Simon Kuznets Kharkiv National University of Economics, Ukraine, e-mail: prokopovichsv@gmail.com

Liubov O. Chagovets, ORCID: 0000-0003-4064-9712, Simon Kuznets Kharkiv National University of Economics, Ukraine, e-mail: liubov.chahovets@hneu.net

МЕТОДОЛОГІЯ АНАЛІЗУ ДАНИХ ЕКОНОМІЧНОГО РОЗВИТКУ КРАЇН / В.В. Донець, В.Є. Стрілець, М.Л. Угрюмов, Д.О. Шевченко, С.В. Прокопович, Л.О. Чаговець

Анотація. Досліджено питання удосконалення методів ідентифікації економічних об'єктів та їх аналізу з використанням алгоритмів інтелектуального оброблення даних. Використання розробленої методології в економічному аналізі дозволяє підвищити якість управління та може бути основою для створення систем підтримання прийняття рішень для попередження потенційно небезпечних змін економічного стану об'єкта дослідження. Запропоновано удосконалений метод кластеризації даних c-середніх з агентно-орієнтованою модифікацією, для визначення відповідності отриманих кластерів актуальним пропонується радіально-базисна нейромережа та її розширення – для аналізу інформативності змінних стану й отримання підмножини інформативних змінних. Розглянуто вплив застосування стиснення даних за допомогою автокодувальника на точність застосування методів. За результатами тестування розробленої методології було доведено зменшення ймовірності неправильного визначення стану під час ідентифікації станів економічних систем та отримано зменшене значення помилки третього роду під час класифікації станів об'єктів.

Ключові слова: машинне навчання, цифровий розвиток, нечітка кластеризація, радіально базисні нейромережі, логістична регресія, аналіз інформативності змінних.

UDC 004.4 + 004.9
DOI: 10.20535/SRIT.2308-8893.2023.4.03

BLOCKCHAIN TRANSACTION ANALYSIS: A COMPREHENSIVE REVIEW OF APPLICATIONS, TASKS AND METHODS

Ya. DOROGYY, V. KOLISNICHENKO

Abstract. Blockchain transaction analysis is a powerful tool to gain insights into the actions and conduct of participants within blockchain networks. This article aims to extensively examine the applications, tasks, and methods associated with blockchain transaction analysis. We look at various uses of transaction analysis, ranging from its instrumental role in blockchain development to its pivotal significance in the field of criminal investigations. By leveraging common techniques and technologies employed in conducting such an analysis, we unlock hidden insights and uncover information that is not visible at first look. This article offers a wide-ranging perspective on the profound significance of blockchain transaction analysis while shedding light on its key role within the cryptocurrency industry and its wide-ranging applications beyond.

Keywords: blockchain transactions, transaction analysis, transaction tracing, flow analysis, blockchain forensics.

INTRODUCTION

Blockchain technology has revolutionized the way financial transactions are conducted and recorded, creating a public decentralized network that eliminates the need for intermediaries and enables secure and transparent transactions. With the growing popularity of cryptocurrencies and blockchain-based systems, the need for effective blockchain transaction analysis has become increasingly important. Blockchain transaction analysis refers to the process of examining and interpreting blockchain data to gain insights into the flow of transactions, identify patterns, and detect anomalies.

This paper provides a review of the applications of analysis in various domains, the methods and techniques used to analyze blockchain data. The paper is organized as follows. First, we provide an overview of blockchain technology and its key concepts including the types of data available on the blockchain. Then, we delve into the applications of blockchain transaction analysis, including cryptocurrency investigations. We provide real-world examples of how blockchain transaction analysis has been used in different domains and discuss the benefits and limitations of the approaches.

Next, we discuss blockchain transaction analysis, the challenges of analyzing blockchain data, and the methods and techniques used to perform blockchain transaction analysis.

Finally, we conclude the paper with a discussion of the future of blockchain transaction analysis, including the challenges and opportunities that lie ahead. We argue that blockchain transaction analysis has the potential to transform many industries by providing greater transparency, security, and efficiency. However, the field is still in its early stages, and much research is needed to develop more effective methods and tools for analyzing blockchain data.

Overall, this paper aims to provide a comprehensive overview of blockchain transaction analysis, covering both the methods and applications of the field. By doing so, we hope to contribute to the growing field of research on blockchain technology and its potential impact on various industries.

BLOCKCHAIN TRANSACTIONS

Blockchain transaction, in simple terms, can be defined as a record of the transfer of digital assets or the storage of information on a blockchain network that is permanently recorded on a distributed ledger.

One of the most notable features of blockchains is that everything stored is visible to everyone, meaning anyone can see who makes transactions to whom. While it may sound easy at first, it appears much more complex.

Mechanisms of asymmetric cryptography are used to define the sender or receiver of a transaction – addresses are formed from public keys, and private keys are used to sign transactions (to prove that the actual owner of the funds created the transaction).

Another concept blockchain networks are using is hierarchical deterministic (HD) wallets. In HD wallets a master seed is used to generate an unlimited number of public-private key pairs, allowing for the creation of multiple addresses and sub-wallets that can be easily managed from a single mnemonic phrase or seed. This enables users to receive and send funds with new addresses each time, therefore increasing the privacy of the end-user.

In terms of record-keeping, there are two common models: unspent transaction output (UTXO) model and account model. In the UTXO model (Fig. 1), each transaction creates a list of outputs that will be spent in future transactions (used as inputs). The outputs are assigned to the addresses that should be able to use (spend) them. The total balance of the address is the sum of all unspent outputs to this address at the current moment.

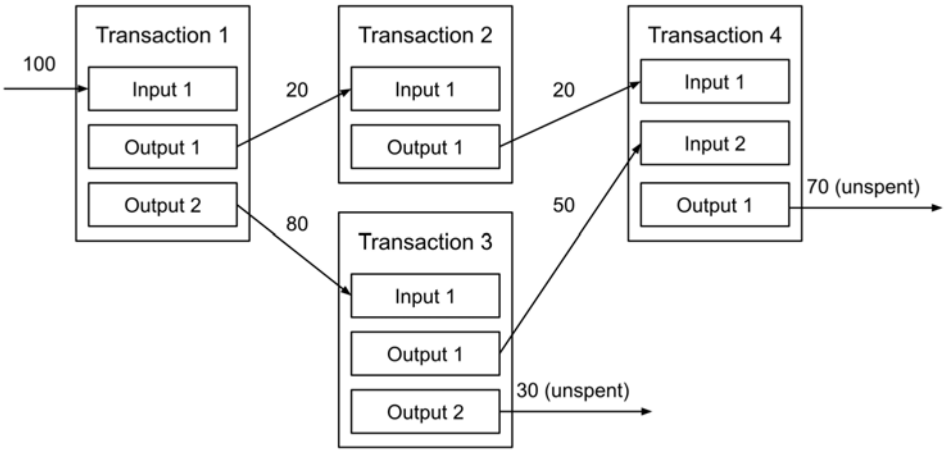


Fig. 1. Simplified UTXO model

The account model, on the other hand, is much simpler to understand. A blockchain maintains balance for each account and keeps a record of all transactions that have affected that balance.

In order to have the ability of multi-user ownership of funds, or more generally, to set up conditions and rules for spending funds (rules of ownership), blockchain networks were built with additional complexity. In fact, Bitcoin transactions do not have sender or receiver fields. Instead, Bitcoin uses lock (scriptPubKey) and unlock (scriptSig) scripts to create a concept of a puzzle, solvable by meeting specified conditions (e.g., to spend a transaction the one should specify a signature in the unlock script, whose public key is set in the lock script).

Bitcoin scripts are extremely limited and do not allow creating complex logic. To face this, Ethereum network uses a concept of smart contract [1], which enables creating complex programs using JavaScript-like language called Solidity, storing them on the blockchain and executing thorough Ethereum Virtual Machine.

Scripting and programming features give extensive possibilities to build applications with various levels of complexity providing end-users with secure decentralized financial (DeFi) services and developers with tools for further optimization (e.g., layer 2 networks) and development [2].

If we talk about what information is stored in the blockchain, then it is usually logical information. By logical information here we mean information related to blocks, transactions, accounts, etc. In other words, data to support the business logic of a blockchain.

There is much information that is not part of blockchain and, usually, it is more technical and does not influence the business logic directly. Let us consider a simplified process of including transactions into the blockchain (which is similar among different networks). A user creates a transaction and signs it, providing proof of address ownership and transaction integrity. After a transaction is signed, the user broadcasts it by using one's own node or through the JSON-RPC interface of a chosen public node. Traveling through a bunch of nodes, the transaction finally reaches miners who include it to the block and mine it. After the block is mined, it gets broadcast to the rest of the nodes. When the rest of the nodes accept it, it is considered as a part of the blockchain. No networking information (IP of the sender and nodes that broadcasted the transaction, etc.) in this process is included into the blockchain, however, intermediate hosts may store it in their own databases.

Taking into consideration all the mentioned specifics, it is not easy to analyze transactions and data stored in the blockchain – who owns the funds, how much, who was the actual sender, what logic the transaction performs, etc. In the next chapter we will go into why such analysis is important and where it is applied.

APPLICATION OF ANALYSIS

Blockchain transaction analysis is a powerful tool that allows us to better understand the behavior of users and events on blockchain networks. As blockchain and decentralized finance (DeFi) technologies continue to grow, there are an increasing number of use cases for transaction analysis. This chapter will explore the key applications of blockchain transaction analysis, including cryptocurrency investigations, risk management, tax compliance, and many others. By leveraging the insights gained through transaction analysis, stakeholders can seize the big picture and make informed decisions.

While some categories may overlap, we think the following distinguishment reflects the unique specifics in the best way.

Crime Investigation

Cryptocurrencies possess unique properties such as decentralization, independence from banks, security, ubiquity, and anonymity. As is the case with other types of assets, these distinct characteristics determine the specific applications of cryptocurrencies. However, these same properties have also made cryptocurrencies an appealing tool for illicit activities such as money laundering, fraud, scam, and sanctions evasion, among others [3; 4].

The Africrypt incident is one of the biggest that has happened with the involvement of cryptocurrencies. Two founders of Africrypt alleged that their firm was hacked resulting in the theft of all its assets. After the statement, the founders vanished. Approximately $3.6 billion in Bitcoin has disappeared in total [5]. As of now, not much added information has been found regarding this case. Law enforcement authorities are reportedly continuing their search for the founders [6]. This incident, among many others [7; 8], is similar to traditional finance scams, where founders (whose names are often known) collect money and disappear.

While blockchain technologies provide a certain level of privacy it cannot be considered as fully anonymous [9], and in many cases a careful analysis may answer the question “where the money goes” [10]. Let us examine some prominent cases where transaction analysis was helpful for the investigation.

Cryptocurrencies are often used as a means of payment in cyber extortion and ransomware attacks. Hackers who carry out these attacks demand payment in cryptocurrency in exchange for returning control of the victim’s computer system or stolen data. One of such cases is NetWalker malware, which is built as Ransomware as a Service (RaaS) model [11], where affiliates rent malware from operators to launch attacks. One of the affiliates was arrested, and a blockchain transaction analysis solution was used to help to track down addresses associated with the affiliate [12].

Hacking of the DeFi projects is quite widespread [13; 14]. Compared to other domains, in the blockchain domain a hacker directly operates the valuable assets such as coins or tokens. It is worth mentioning that the biggest amounts of assets are concentrated in cross-chain bridges and centralized crypto exchanges (CEX) which make them attractive targets [15]. Transaction analysis is usually applied to get an understanding of the attack and to track the ones who were involved.

An attack analysis is an essential measure to be taken after the incident has occurred. This process involves identifying how the system was compromised, assessing the extent of the damage caused, determining strategies for minimizing the damage, and addressing any vulnerabilities that were exploited. To identify how the blockchain system was hacked, attackers’ transactions together with involved smart contracts are analyzed. Such analysis is often performed by the owning company, investigating company or blockchain community [16; 17] with various levels of details.

Hackers who steal funds from blockchains often seek to launder the stolen cryptocurrencies to conceal their identities and make it difficult for law enforcement agencies to trace the illicit funds [18]. They can do this by using mixers, tumblers, or other obfuscation techniques to obscure the trail of transactions and make it hard to trace the stolen funds. Additionally, hackers can use decentralized exchanges to convert stolen cryptocurrencies to other assets, such as privacy coins or stablecoins, to further obfuscate the trail of transactions. These assets can then be moved through multiple wallets to further distance the funds from the original theft. The final step may involve converting the stolen cryptocurrencies to fiat currency through a regulated exchange or other means to cash out the illicit funds.

One of the successful investigations of laundering is the Bitfinex case. According to Elliptic [19], after the hack stolen funds were slowly being laundered using different techniques. AlphaBay is one of the services that was used as a mixer to hide the trails. However, later it was seized by law enforcement, and this likely allowed them to get trails to the hackers.

Another, less successful investigation case, is a hack of Zaif exchange in September 2018. Crystal Blockchain Analytics engineering team performed an analysis of bitcoin movements [20] and could find addresses involved in the hack. Although the owners of the addresses are unknown, the addresses are being monitored in case of further transactions.

One notable type of blockchain assets that got much attention is Non-Fungible Token (NFT). NFTs are assets that represent ownership of unique items such as music, videos, art or other on a blockchain network. They are not dividable and interchangeable with one another. Each NFT is unique and cannot be replicated. Although it can be implicated in criminal activities similar to other digital assets, one distinctive aspect worth mentioning is copyright infringement. Blockchain technology can guarantee the uniqueness of the token but cannot guarantee the uniqueness of the represented asset, which can be copied. Transaction analysis can be used to assist with NFT copyrighting. By analyzing the NFT transactions it may be possible to verify its authenticity and identify the original creator or owner of the work. This information could also be used as evidence to support copyright claims. One notable project that tries to detect copyright infringement by scanning blockchains and marketplaces is DeviantArt [21]. They use different techniques including machine learning to spot the copy.

We can observe that blockchain transaction analysis is used as a valuable tool in crime investigations, enabling law enforcement agencies to track the flow of funds in the blockchain network and identify any suspicious activity associated with illegal activities such as money laundering, dark web transactions, cybercrime, and fraud. Transaction analysis also helps trace the flow of funds associated with cyberattacks, ransomware payments, and other malicious activities. It provides insights into transaction behavior and patterns that can be used to identify potential criminal activity and take appropriate action.

Compliance and Regulation

Cryptocurrency regulations are laws or rules established by governments or regulatory bodies to govern the use, trading, and custody of cryptocurrencies. These regulations aim to protect investors, prevent illicit activities such as money laundering and terrorist financing, and promote the stability and integrity of the financial system. Cryptocurrency regulations can cover a wide range of topics, depending on the jurisdiction and the specific concerns of regulators [22]. While these regulations are mostly related to cryptocurrencies rather than technology itself, some countries may try to enforce regulations on the tech side too (e.g., on mining) [23].

Transaction analysis is a useful tool for enforcing cryptocurrency regulations and ensuring compliance with regulatory requirements [24; 25]. Regulators can use transaction analysis to monitor and detect potential money laundering activities, enforce KYC (know your customer) and tax compliance, prevent fraud, and protect consumers in the cryptocurrency market [26].

By analyzing transaction patterns and identifying any unusual or suspicious activity, regulators can take appropriate action to prevent money laundering and other financial crimes. They can also use transaction analysis to monitor compliance with know-your-customer requirements and tax laws and regulations. Additionally, transaction analysis can help prevent cryptocurrency fraud by identifying any fraudulent activity and taking appropriate action.

While in the previous section transaction analysis is applied after an event happened (for the investigation), in case of regulations, transaction analysis is mostly used continuously (for the detection and prevention).

Trade and investment

In traditional finance, financial transactions are mostly opaque, and investors often rely on intermediaries [27] to provide information about the assets they are investing in. Investors and traders use methods such as technical analysis to analyze financial markets and securities based on statistical trends and patterns in historical price and volume data. Trading in blockchain offers new opportunities and challenges with its unique characteristics of transparency, security, and decentralization [28].

Having access to transaction data changes the rules of the game. However, without a proper processing of massive amounts of raw data transparency does not give you advantages. Therefore, it is important to produce new methods and tools that can provide insights into the behavior of market participants and the underlying fundamentals of digital assets. Moreover, these methods and tools should be the same or better than in your potential opponents, as they also have access to the same raw data.

Here transaction analysis can be helpful in several ways. It can provide volume and velocity of transactions for a particular cryptocurrency [29]. Blockchain transaction analysis can give information on the distribution of digital assets among market participants and provide valuable insights into their behavior. This information can help traders and investors identify potential price levels for a particular cryptocurrency based on the level of demand from buyers and sellers. By analyzing this information traders and investors can gain insights into participants' trading strategies and use this information to adjust their own trading decisions.

Besides that, it can be used for analyzing the flow of assets within a blockchain to identify large transactions and movements of funds that may be indicative of market manipulation or other illicit activities. Transaction flow analysis can help traders avoid entering into positions that may be vulnerable to sudden price movements.

Risk Management

Organizations that try to adopt blockchain and DeFi technologies for their businesses should be aware of numerous additional risks [30–32].

Cryptocurrencies are still growing and one of the primary risks is unclear regulations, specifically, legal, and regulatory compliance. Blockchain-based businesses may face challenges in complying with current regulations or in predicting future regulations, which can result in legal and financial penalties or reputational damage. The risk of unclear regulations in blockchain risk management is significant because blockchain technology operates in a regulatory gray area in many countries.

Another set of risks comes from the technical side. Bugs, vulnerabilities, network scalability difficulties can lead to various negative outcomes such as loss or theft of funds [33], network downtime [34], reputational damage and others.

Volatility and liquidity are another two significant risks associated with blockchain and cryptocurrencies [35]. These risks can affect both investors and businesses that use cryptocurrencies for transactions or other purposes. Volatility can lead to significant losses for investors who have invested in cryptocurrencies, as the value of their investments can decrease rapidly. Additionally, businesses that use cryptocurrencies for transactions can be negatively affected by volatility as the value of their transactions can also fluctuate rapidly. Cryptocurrency markets can be relatively illiquid, particularly for less popular cryptocurrencies or during periods of market instability. This illiquidity can make it difficult for investors to sell their cryptocurrencies when they need to, leading to losses. Additionally, illiquidity can create challenges for businesses that use cryptocurrencies for transactions, as it can be difficult to find a buyer or seller for the desired cryptocurrency at a fair market price.

Transaction analysis is a useful tool for risk management in blockchain. It can provide businesses with insights into transaction behavior and patterns, which can be used to identify potential risks and vulnerabilities. Transaction analysis can be used to detect fraudulent activity, such as money laundering or other financial crimes, by analyzing transaction patterns and identifying any unusual or suspicious activity. It can also help businesses monitor compliance with regulations and industry standards by detecting any potential compliance issues. Businesses can determine the level of risk associated with a particular transaction or customer and take appropriate action to manage that risk. Moreover, transaction analysis can help investors and businesses assess the liquidity of cryptocurrencies by analyzing transaction volumes.

Supply Chain Management

Supply chain management in blockchain refers to the use of blockchain technology to track and manage the movement of goods and services through a supply chain. Blockchain technologies provide a transparent and secure platform for tracking and verifying transactions in real-time [36]. Each asset is represented through a unique token. When a party performs transfer of the asset, it also creates and signs a transaction to transfer the token (that represents actual asset) on that blockchain. Transactions are then recorded on the blockchain, and the entire process is transparent for the shareholders. This can help businesses to optimize their supply chain operations, reduce costs, and ensure compliance with relevant regulations and industry standards.

In the supply chain process, blockchain transaction analysis is a core tool, which allows stakeholders to follow the entire process. It allows extracting and analyzing transaction patterns, businesses can gain valuable insights into the movement of goods and services through the supply chain [37–39].

Blockchain Development

Analysis of transactions is also important for blockchain development and its optimization. At different stages of development transactions are analyzed to debug errors [40] and monitor the network health. It is used to get understanding about users' behavior inside the network, to identify their needs and troubles [41]. By analyzing transaction patterns, developers can identify bottlenecks in the blockchain network, such as congested nodes [42] or high transaction fees. The information gained can be used to develop new solutions to optimize the blockchain platform. Such optimization may apply to its performance [43; 44] or security [45]. Additionally, transaction analysis allows developers to detect suspicious activity or DoS attacks and take steps to mitigate the risks [46].

Blockchain Attacks Detection and Prevention

Real-time analysis of transactions is employed for monitoring smart contracts to detect possible attacks and prevent them. The analysis of transaction data in real-time enables the detection of any suspicious activity, allowing for timely intervention to prevent or minimize the impact of an attack.

There are no strict criteria defining what to consider as an attack, therefore various heuristics (detect maximum value transfer, ownership change, contract upgrades) and machine learning algorithms for flow analysis may be used. When a potential attack happens and the algorithm detects it, the stakeholders get notified so they can perform further actions. In situations where immediate response is required, it is possible to configure automated actions, such as temporarily halting the core functionality of a contract.

One such widely used solution is Forta [47]. It gets advantage of transaction analysis to detect and mitigate security threats in decentralized applications and smart contracts. Forta technology is designed to analyze blockchain transactions and data to identify and prevent hacks, exploitations, and other malicious activities. It is stated [48] that the utilization of the system could have prevented numerous attacks and financial losses.

TASKS AND METHODS

Given the applications of the analysis described in the previous sections, we have identified and selected the most frequent and critical tasks to be addressed through blockchain transaction analysis, which can be broadly categorized into three big groups. In this section, we will examine these tasks and the techniques employed to address them.

Linking addresses with identities

A common task in transaction analysis is to identify the owner of the address. By owning an address, we mean that the person holds a private key (or seed/mnemonic phrase) and a corresponding public key, from which the address is created. Because addresses are created using solely cryptographic mechanisms and even before interacting with the chain, it can be impossible to get the real identity of the owner. Fortunately, we do not need to solve the problem where users just create their addresses, but where users actually use them. Similar to this task, there is an opposite one – to find addresses belonging to a certain entity.

One of the most straightforward methods for linking addresses to identities is to require users to disclose their identities when they purchase or sell cryptocurrency with fiat money. This is a prevalent regulatory approach, and most cryptocurrency exchanges now must follow Know Your Customer (KYC) procedures that include several steps to identify the user. KYC procedure typically involves several steps, such as providing identification documents and verifying the user's personal information, in order to confirm the user's identity. Once a user has been successfully identified through this process, their cryptocurrency transactions on the exchange can be linked to their real-world identity, making it easier to track any suspicious activity or money laundering attempts.

When a signed transaction or block is transmitted to other nodes, or a JSON-RPC call is made on a public network, information about the sender, such as their IP address, can be recorded by the nodes and intermediary network devices. This can provide a means of identifying the user in the future. In general, transmitting any information related to a blockchain address to a third-party server, such as making a purchase on a website, searching for a transaction, or checking a balance on a blockchain explorer [49], or using a wallet application that utilizes analytics, can potentially establish a connection between the user and the address.

Another method for mapping entities to addresses is to maintain a record of information related to the blockchain addresses that has been published by the entity or utilize openly accessible databases. One example of such a database is a list of malicious actors [50] or a database of sanctioned addresses, which can be employed during analysis. Social networks scraping may also be helpful, as users often publish addresses of their wallets. The main downside of this approach is you need to set up complex infrastructure and collect a lot of data beforehand, and the identity you are interested in may not be even in this data.

Flow Comprehension

By blockchain transaction flow we refer to sequences of transactions that occur on a blockchain network. These sequences can vary from small to large and have complex structures containing branches and joins. Complex structures may contain valuable information that is not visible at first look and therefore different methods should be applied to extract it.

Occasionally, these sequences can intentionally have intricate structures. Criminal actors often obscure traffic, expecting that investigators will lose track. However, by utilizing the right approaches and tools, it is possible to gain significant insights into the flow and uncover details that may have otherwise remained hidden. In the following sections, we will explore common approaches to analyzing blockchain transaction flows.

One can manually retrieve data from the blockchain using blockchain explorers or similar tools that enable communication with blockchain nodes. Usually, they are web-based tools [51] that enable users to access and navigate the contents of a blockchain. They have a graphical interface to examine and analyze blockchain data, such as transaction records, account addresses, and balances. The primary function of a blockchain explorer is to facilitate the search of specific transactions, verify wallet balances, and examine network metrics. Although blockchain explorers are useful for basic cases, they are not suitable for handling complex cases involving long chains of transactions.

Graph visualization is used to handle the complexity of sequences of transactions. Different kinds of graphs and representations can be used to fit certain needs [52–54]. However, in the majority of cases, it is desirable to present the flow as a graph, in which addresses are represented as nodes and transactions as directed edges (Fig. 2). This format gives an ability to follow the asset transfer in the most natural way. It can still be challenging to comprehend, and as such, it is beneficial to group, filter, highlight, and conceal distinct elements, to extract or segregate valuable information from irrelevant data.

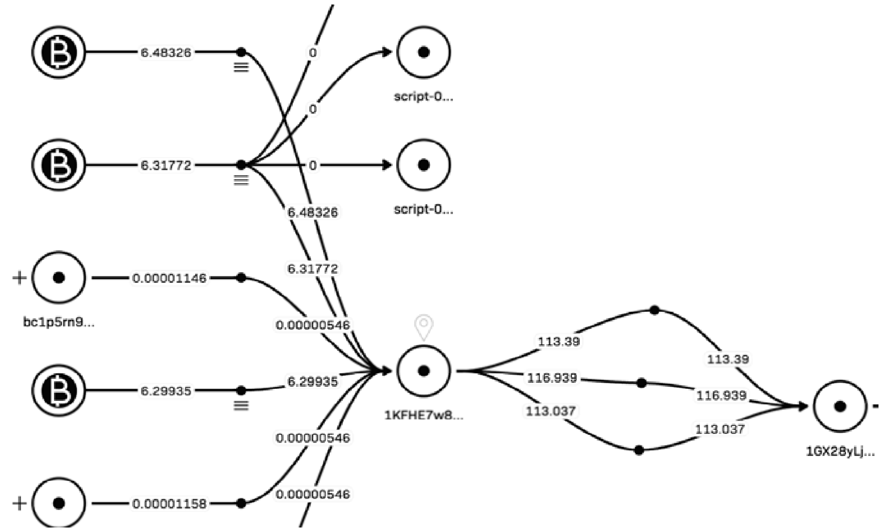


Fig. 2. Graph visualization in Crystal Explorer [55]

For various reasons, users may want to transfer their assets from one network into another or want to exchange one type of assent into another. They use exchange platforms and cross-chain bridges for these purposes. Bad actors who want to obfuscate their traffic can also take advantage of these methods. Additionally, they can use tools such as mixers [56; 57] which can significantly complicate the task of investigators attempting to comprehend the transactional flow. We can define these instruments as conversion protocols. Transactions tracing tools should be aware of different conversion protocols and be robust enough to perform address linking (in cases where theoretically possible). In many cases conversion protocols, such as cross-chain bridges [58], are well-documented and when they are not specifically designed for hiding traffic it can be an easy task to find what is the address of the user in the other network. In cases where no documentation is available or where it is claimed that the system provides absolute anonymity, it may be necessary to perform a manual analysis of the system. A thorough manual analysis of the system can provide valuable insights into its functionality. It helps in understanding the meaning of user transactions and identifying any potential flaws that could be exploited. Additionally, this type of analysis can reveal possibilities of developing new methods to obtain additional information about the transactions or the users [59].

In order to simplify flow analysis, various algorithms can be utilized to discover connections, patterns, and anomalies [60]. They can be used to simplify the view or bring the most important data to the front. These algorithms can be a classic one [61] or machine learning algorithms [62].

There are commercial tools available on the market, such as Chainalysis Reactor, MistTrack and others [63], that provide convenient instruments for flow analysis, including graph visualizations and various other features discussed in this section.

Smart Contract Brakedown

In contrast to the regular banking transactions, blockchain transactions became more than just funds transferring. Smart contracts let writing very complex conditions for transferring funds, and as a result – building additional abstraction layers and protocols. This allowed the creation of a new financial paradigm – DeFi.

With increased complexity, transaction analysis became harder and more time consuming to perform. Calling a certain function on a smart contract and transferring funds can mean different things and therefore prior contract understanding is needed. A call to a smart contract may create a chain of calls with different arguments, including calls to other contracts. A list of methods has been developed to approach the smart contract breakdown.

Manual source code review provides a comprehensive insight into the behavior of a smart contract. This process involves a thorough examination of the code, line by line, to gain understanding of both the overarching concept and the finer points. However, this type of analysis requires extensive knowledge of programming languages, cryptography, and blockchain technology, and can be a time-consuming process. Reading documentation of a product may be helpful and can clarify reasons behind some programming decisions or explain unfamiliar concepts. However, it is not always available.

During a source code review, the availability of the source code is another important aspect to consider. It is common for smart contract source code to be published and, in many cases, it can be found on GitHub. However, having a source code of the contract does not mean the exact same contract is published on the blockchain, therefore a contract verification is needed [64] – to match source code to on-chain bytecode.

In some cases, developers may choose not to make the source code of the contracts publicly available. As a result, alternative techniques are necessary to gain insight into the behavior of the contract through analyzing its bytecode. Generally, the process is called reverse engineering. It is similar to the code review but more complex and requires more effort. The reason for this is that a compiled smart contract contains significantly less information compared to the original code. In cases of optimizations the resulting bytecode gets even more complicated, as human written code is converted into more efficient code patterns. To simplify the reverse engineering process disassemblers (convert bytecode to EVM opcodes) and decompilers are used [65; 66]. Decompilers convert a bytecode to high-level representation. However, due to the loss of information during compilation, the code does not look like the original code.

If we want to look at the actual execution of the smart contract on the chain, block explorers may be helpful for simple cases. Some of them have transaction decoders and can provide execution traces. There are tools developed specifically for transaction decoding, such as Transaction Tracer [67] or similar [68], which provide a call trace, which is a tree of function calls and arguments, made through different contracts during transaction execution. Furthermore, there are tools for local EVM tracing [69], which allow detailed examination of smart contract transactions. Development environments, like Truffle, have even more convenient means to debug on-chain transactions [70].

Automated analysis tools for smart contracts can be used to get a better understanding of smart contracts. Usually, they are divided into two categories – static and dynamic analyzers [71].

Static analysis tools perform contract analysis without running them. Slither framework [72], is one of such tools, is designed to automatically find vulnerabilities, give information about the contract and its functions, give summary about the authorization accesses and many other.

Dynamic analysis tools, on the other hand, perform analysis by executing smart contracts or their parts. Various classes of dynamic analysis tools used for analyzing smart contracts such as symbolic execution tools, Satisfiability Modulo Theories (SMT) solvers, taint analyzers and fuzzers [73]. Mythril, Echidna [74] and Manticore [75] are one of the most widely used tools to find vulnerabilities in the code, to find a set of inputs that transit a program into an unexpected state or to explore all possible states. These tools and approaches are not mutually exclusive, but rather they give different perspectives on how a smart contract works.

Commercial tools like MythX [76] combine static and dynamic approaches to get the best of both worlds and provide most comprehensive results.

Recent research and developments in artificial intelligence (more precisely, large language models such as ChatGPT [77]), allowed using these technologies for explaining the code, reverse-engineering [78] and even for finding vulnerabilities [79]. These tools are already used now and will be even more adopted in the near future to assist during the code analysis.

CONCLUSIONS

Blockchain transaction analysis is a crucial tool for gaining insights into the behavior of users on blockchain networks. From anti-money laundering and fraud detection to supply chain management and tax compliance, there are many applications for transaction analysis in the world of cryptocurrency and beyond.

Despite the challenges posed by the anonymous and decentralized nature of blockchain networks, there is a growing awareness of the importance of transparency and accountability in the cryptocurrency industry. By utilizing the insights gained through transaction analysis, regulators, businesses, and other stakeholders can work together to build a more secure, efficient, and sustainable blockchain ecosystem.

The methods and techniques used in transaction analysis continue to evolve, and there is a growing need for more sophisticated tools to keep pace with the complexity of blockchain networks. Advances in machine learning, graph analysis, and other data science techniques are likely to have a significant impact on the future of blockchain transaction analysis.

In our future work we will analyze multiple blockchain networks to get advantages of their specifics to improve and develop new methods for analyzing transactions. We will dive into protocols at different layers and develop solutions to extract additional information that is not available using traditional methods.

REFERENCES

1. V. Buterin, “Ethereum: A next-generation smart contract and decentralized application platform,” *Ethereum.org*. Accessed on: April 24, 2023. [Online]. Available: https://ethereum.org/669c9e2e2027310b6b3cdce6e1c52962/Ethereum_Whitepaper_-_Buterin_2014.pdf
2. S. Sharma and M. Naggar, “A New Era for Bitcoin?,” *Binance Research*. Accessed on: April 24, 2023. [Online]. Available: https://research.binance.com/static/pdf/a-new-era-for-bitcoin.pdf
3. “The 2023 Crypto Crime Report,” *Chainalysis.com*, 2023. Accessed on: April 24, 2023. [Online]. Available: https://go.chainalysis.com/rs/503-FAP-074/images/Crypto_Crime_Report_2023.pdf
4. “Blockchain Security and AML Analysis Report - 2022 Annual,” *Slowmist.com*. Accessed on: April 24, 2023. [Online]. Available: https://www.slowmist.com/ report/2022-Blockchain-Security-and-AML-Analysis-Annual-Report(EN).pdf
5. L. Prinsloo and R. Henderson, “Trail of Brothers Linked to Missing Bitcoin Stash Is Still Murky,” *Bloomberg.com*. Accessed on: April 24, 2023. [Online]. Available: https://www.bloomberg.com/news/articles/2021-06-27/trail-of-brothers-linked-to-missing-bitcoin-stash-is-still-murky
6. L. Prinsloo, “Crypto Losses Probed by South African Cops After Brothers Vanish,” *Bloomberg.com.* Accessed on: April 24, 2023. [Online]. Available: https://www.bloomberg.com/news/articles/2022-01-11/crypto-losses-probed-by-south-african-cops-after-brothers-vanish
7. “Top five most wanted crypto criminals,” *CNBCTV18*. Accessed on: April 24, 2023. [Online]. Available: https://www.cnbctv18.com/cryptocurrency/top-five-most-wanted-crypto-criminals-15897891.htm
8. “The 10 biggest crypto scams on record and the lessons we can learn from them,” *Irishtechnews.ie.* Accessed on: April 24, 2023. [Online]. Available: https://irishtechnews.ie/10biggestcryptoscams/
9. “Protect your privacy,” *Bitcoin.org.* Accessed on: April 24, 2023. [Online]. Available: https://bitcoin.org/en/protect-your-privacy
10. “Ukrainian Cyber Police Department in Collaboration with Crystal,” *Crystalblockchain.com.* Accessed on: April 24, 2023. [Online]. Available: https://crystalblockchain.com/articles/ukrainian-cyber-police-department-now-in-collaboration-with-crystal-blockchain/
11. K. Baker, “What is Ransomware as a Service (RaaS)?,” *crowdstrike.com*. Accessed on: April 24, 2023. [Online]. Available: https://www.crowdstrike.com/cybersecurity-101/ransomware/ransomware-as-a-service-raas/
12. “Chainalysis in action: U.S. authorities disrupt NetWalker ransomware,” *Chainalysis*. Accessed on: April 24, 2023. [Online]. Available: https://blog.chainalysis.com/ reports/netwalker-ransomware-disruption-arrest/
13. “Rekt - leaderboard,” *rekt.* Accessed on: April 24, 2023. [Online]. Available: https://rekt.news/leaderboard/
14. “SlowMist Hacked - SlowMist Zone,” *Slowmist.io.* Accessed on: April 24, 2023. [Online]. Available: https://hacked.slowmist.io/
15. “The 10 biggest crypto exchange hacks in history,” *Crystalblockchain.com.* Accessed on: April 24, 2023. [Online]. Available: https://crystalblockchain.com/ articles/the-10-biggest-crypto-exchange-hacks-in-history/
16. @officer_cia, “How cross-chain bridges are hacked? A detailed review,” *Mirror.xyz*. Accessed on: April 24, 2023. [Online]. Available: https://officercia.mirror.xyz/ AFkEUuxid1egNm4XdqYEzWEwosPNbz2CNghlNrq7LZQ
17. @officer_cia, “Retrospective: hacks in web3,” *Telegraph*. Accessed on: April 24, 2023. [Online]. Available: https://telegra.ph/Retrospective-hacks-in-web3-10-24

18. SlowMist, "SlowMist AML: Tracking funds laundered by Tornado Cash," *Medium*, Accessed on: April 24, 2023. [Online]. Available: https://slowmist.medium.com/ slowmist-aml-tracking-funds-laundered-by-tornado-cash-3a0e1f637054
19. "New York husband and wife arrested for laundering bitcoin," *Elliptic.co.* Accessed on: April 24, 2023. [Online]. Available: https://www.elliptic.co/blog/elliptic-analysis-new-york-husband-and-wife-arrested-for-laundering-5-billion-in-bitcoin-stolen-from-bitfinex-in-2016
20. "Crystal Blockchain Analytics: Investigation of the Zaif Exchange Hack," *Bitfury.com.* Accessed on: April 24, 2023. [Online]. Available: https://bitfury.com/ content/downloads/bitfury_crystal_zaif_report_23_10_18.pdf
21. "DeviantArt protect: Helping safeguard your art," *Deviantart.com*. Accessed on: April 24, 2023. [Online]. Available: https://www.deviantart.com/team/journal/ DeviantArt-Protect-Helping-Safeguard-Your-Art-884278903
22. "Cryptocurrency regulations around the world," *ComplyAdvantage*. Accessed on: April 24, 2023. [Online]. Available: https://complyadvantage.com/insights/ cryptocurrency- regulations-around-world/
23. T. Akhtar and S. Shukla, "China Makes a Comeback in Bitcoin Mining Despite Government Ban," *Bloomberg.com*. Accessed on: April 24, 2023. [Online]. Available: https://www.bloomberg.com/news/articles/2022-05-17/china-makes-a-comeback-in-bitcoin-mining-despite-government-ban
24. "Cryptocurrency transaction monitoring: What you need to know," *ComplyAdvantage*. Accessed on: April 24, 2023. [Online]. Available: https://complyadvantage. com/insights/transaction-monitoring-cryptocurrencies/
25. "How continuous cryptocurrency transaction monitoring gives compliance teams peace of mind," *Chainalysis*. Accessed on: April 24, 2023. [Online]. Available: https://blog.chainalysis.com/reports/kyt-continuous-monitoring/
26. "Cryptocurrency regulation: How governments around the world regulate crypto," *Chainalysis*. Accessed on: April 24, 2023. [Online]. Available: https://blog.chainalysis.com/reports/cryptocurrency-regulation-explained/
27. N. Asokan, "Financial Intermediaries: their role on real examples," *Agicap.com*. Accessed on: April 24, 2023. [Online]. Available: https://agicap.com/en/article/ financial-intermediaries/
28. M. Morel, "Technical analysis is dead, long live transaction analysis," *CoinDesk*. Accessed on: April 24, 2023. [Online]. Available: https://www.coindesk.com/ layer2/2022/10/26/technical-analysis-is-dead-long-live-transaction-analysis/
29. I.G.A. Pernice, G. Gentzen, and H. Elendner, "Cryptocurrencies and the Velocity of Money," *Cryptoeconomic Systems*, vol. 1, iss. 1, 2021. doi: 10.21428/ 58320208.f212c00e.
30. U.S. Department of the Treasury, "Illicit Finance Risk Assessment of Decentralized Finance," *Treasury.gov.* Accessed on: April 24, 2023. [Online]. Available: https://home.treasury.gov/system/files/136/DeFi-Risk-Full-Review.pdf
31. "Cryptocurrency: Risk management overview," *Wtwco.com*. Accessed on: April 24, 2023. [Online]. Available: https://www.wtwco.com/-/media/WTW/Insights/2019/01/ cryptocurrency-risk-management-overview.pdf
32. "Blockchain risk management," *Deloitte.com*. Accessed on: April 24, 2023. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/ financial-services/us-fsi-blockchain-risk-management.pdf
33. M. White, "Web3 is Going Just Great," *Web3isgoinggreat.com*. Accessed on: April 24, 2023. [Online]. Available: https://web3isgoinggreat.com/
34. S. Kessler and D. Nelson, "Polygon blockchain nodes briefly went out of sync, affecting explorer, sowing confusion," *CoinDesk*. Accessed on: April 24, 2023. [Online]. Available: https://www.coindesk.com/tech/2023/02/22/polygon-blockchain-suffers-apparent-outage/
35. T. Chang, J. Ho, Z. Tirrell, G. Weng, and J. You, "A risk classification framework for decentralized finance protocols," *Soa.org*. Accessed on: April 24, 2023. [Online].

Available: https://www.soa.org/4aa5bb/globalassets/assets/files/resources/research-report/2022/decentralized-finance-protocols.pdf

36. V. Gaur and A. Gaiha, "Building a Transparent Supply Chain," *Harvard business review*, 2020.
37. P. Dutta, T.-M. Choi, S. Somani, and R. Butala, "Blockchain technology in supply chain operations: Applications, challenges and research opportunities," *Transportation Research Part E: Logistics and Transportation Review*, vol. 142, Elsevier BV, p. 102067, Oct. 2020. doi: 10.1016/j.tre.2020.102067.
38. R. Pratik, "How AI and Blockchain transforming supply chain management?," *Intuz.com*. Accessed on: April 24, 2023. [Online]. Available: https://www.intuz.com/blog/ai-and-blockchain-in-supply-chain-management
39. .L. Compagnucci, D. Lepore, F. Spigarelli, E. Frontoni, M. Baldi, and L. Di Berardino, "Uncovering the potential of blockchain in the agri-food supply chain: An interdisciplinary case study," *Journal of Engineering and Technology Management*, vol. 65, Elsevier BV, p. 101700, Jul. 2022. doi: 10.1016/j.jengtecman.2022.101700.
40. "Incident report: Rootstock peg-out service outage (Fixed)," *Rsk.com*. Accessed on: April 24, 2023. [Online]. Available: https://blog.rsk.co/noticia/incident-report-rsk-peg-out-service-outage/
41. "Data and analytics," *ethereum.org*. Accessed on: April 24, 2023. [Online]. Available: https://ethereum.org/en/developers/docs/data-and-analytics/
42. "Network congestion," *Bybit Learn*. Accessed on: April 24, 2023. [Online]. Available: https://learn.bybit.com/glossary/definition-network-congestion/
43. "Scaling," *ethereum.org*. Accessed on: April 24, 2023. [Online]. Available: https://ethereum.org/en/developers/docs/scaling/
44. S.D. Lerner, "RSK scalability," *Innovation Stories*. Accessed on: April 24, 2023. [Online]. Available: https://medium.com/iovlabs-innovation-stories/rsk-scalability-c44252f05a4b
45. "State of research: increasing censorship resistance of transactions under proposer/builder separation (PBS)," *HackMD*. Accessed on: April 24, 2023. [Online]. Available: https://notes.ethereum.org/@vbuterin/pbs_censorship_resistance
46. R. Behnke, "How blockchain DDoS attacks work," *Halborn*. Accessed on: April 24, 2023. [Online]. Available: https://www.halborn.com/blog/post/how-blockchain-ddos-attacks-work.
47. "Forta: a decentralized runtime security solution for automated threat detection and prevention on smart contracts," *Forta.network*. Accessed on: April 24, 2023. [Online]. Available: https://docs.forta.network/en/latest/2022-7-11%20Forta%20Litepaper.pdf
48. Forta, "How to use Forta's Threat Intel Data," *Notion*. Accessed on: April 24, 2023. [Online]. Available: https://forta.notion.site/How-Forta-alerted-on-past-hacks-71e63d933ef5426d92642a8019708d48
49. D. Nelson and M. Hochstein, "Leaked slides show how Chainalysis flags crypto suspects for cops," *CoinDesk*. Accessed on: April 24, 2023. [Online]. Available: https://www.coindesk.com/business/2021/09/21/leaked-slides-show-how-chainalysis-flags-crypto-suspects-for-cops/
50. "Bitcoin abuse database," *Bitcoinabuse.com*. Accessed on: April 24, 2023. [Online]. Available: https://www.bitcoinabuse.com/
51. *Etherscan.io*. Accessed on: April 24, 2023. [Online]. Available: https://etherscan.io/
52. N. Tovanich, N. Heulot, J.-D. Fekete, and P. Isenberg, "Visualization of Blockchain Data: A Systematic Review," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 7, pp. 3135–3152, Jul. 01, 2021. doi: 10.1109/tvcg.2019.2963018.
53. J.S. Tharani, E.Y.A. Charles, Z. Hou, M. Palaniswami, and V. Muthukkumarasamy, "Graph Based Visualisation Techniques for Analysis of Blockchain Transactions," *2021 IEEE 46th Conference on Local Computer Networks (LCN)*, Oct. 04, 2021. doi: 10.1109/lcn52139.2021.9524878.

54. "The ultimate guide to graph visualization," *Cambridge-intelligence.com*. Accessed on: April 24, 2023. [Online]. Available: https://info.cambridge-intelligence.com/graph-visualization-white-paper
55. "Crystal," *Crystalblockchain.com*. Accessed on: April 24, 2023. [Online]. Available: https://explorer.crystalblockchain.com/
56. "Introducing cross-Chain Investigations to Reactor," *Chainalysis*. Accessed on: April 24, 2023. [Online]. Available: https://blog.chainalysis.com/reports/cross-chain-investigations/
57. "How cryptomixers allow cybercriminals to clean their ransoms," *Intel471*. Accessed on: April 24, 2023. [Online]. Available: https://intel471.com/blog/ cryptomixers-ransomware
58. "Blockchain bridges: An industry overview," *Rsk.com.* Accessed on: April 24, 2023. [Online]. Available: https://blog.rsk.co/noticia/blockchain-bridges-an-industry-overview/
59. T. Tironsakkul, M. Maarek, A. Eross, and M. Just, "Tracking Mixed Bitcoins," *arXiv*, 2020. doi: 10.48550/ARXIV.2009.14007.
60. M. J. Shayegan and H. R. Sabor, "A Collective Anomaly Detection Method Over Bitcoin Network." arXiv, 2021. doi: 10.48550/ARXIV.2107.00925.
61. Z. Wu, J. Liu, J. Wu, Z. Zheng, and T. Chen, "TRacer: Scalable Graph-based Transaction Tracing for Account-based Blockchain Trading Systems," IEEE Transactions on Information Forensics and Security. Institute of Electrical and Electronics Engineers (IEEE), pp. 1–1, 2023. doi: 10.1109/tifs.2023.3266162.
62. J. Siegenthaler, *Blockchain Clustering with Machine Learning*. Switzerland: University of Basel, 2020.
63. "Top 32 blockchain analysis tools," *Startup Stash*. Accessed on: April 24, 2023. [Online]. Available: https://startupstash.com/blockchain-analysis-tools/
64. "Verifying smart contracts," *ethereum.org*. Accessed on: April 24, 2023. [Online]. Available: https://ethereum.org/en/developers/docs/smart-contracts/verifying/
65. "Bytecode Decompilation," *Contract Library*. Accessed on: April 24, 2023. [Online]. Available: https://library.dedaub.com/decompile
66. "Online Solidity Decompiler," *Online Solidity Decompiler*. Accessed on: April 24, 2023. [Online]. Available: https://ethervm.io/decompile/
67. *OpenChain Monorepo*. Accessed on: April 24, 2023. [Online]. Available: https://github.com/openchainxyz/openchain-monorepo
68. @w1nt3r_eth, "A list of power tools (and their hidden features) that security researchers use to investigate hacks," *Twitter*. Accessed on: April 24, 2023. [Online]. Available: https://twitter.com/w1nt3r_eth/status/1597998923226177543
69. "EVM tracing," *go-ethereum*. Accessed on: April 24, 2023. [Online]. Available: https://geth.ethereum.org/docs/developers/evm-tracing
70. "Announcing our fully featured, portable solidity debugger," *Trufflesuite.com*. Accessed on: April 24, 2023. [Online]. Available: https://trufflesuite.com/blog/announcing-full-portable-solidity-debugger/
71. ConsenSys Diligence, "Static and dynamic analysis - ethereum smart contract best practices," *Github.io*. Accessed on: April 24, 2023. [Online]. Available: https://consensys.github.io/smart-contract-best-practices/security-tools/static-and-dynamic-analysis/
72. J. Feist, G. Grieco, and A. Groce, "Slither: A Static Analysis Framework for Smart Contracts," *2019 IEEE/ACM 2nd International Workshop on Emerging Trends in Software Engineering for Blockchain (WETSEB)*. IEEE, May 2019. doi: 10.1109/wetseb.2019.00008.
73. "Fuzzing," *ConsenSys Diligence*. Accessed on: April 24, 2023. [Online]. Available: https://consensys.net/diligence/fuzzing/

74. G. Grieco, W. Song, A. Cygan, J. Feist, and A. Groce, “Echidna: effective, usable, and fast fuzzing for smart contracts,” *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*. ACM, Jul. 18, 2020. doi: 10.1145/3395363.3404366.
75. M. Mossberg et al., “Manticore: A User-Friendly Symbolic Execution Framework for Binaries and Smart Contracts,” *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, Nov. 2019. doi: 10.1109/ase.2019.00133.
76. “MythX: Preparing for a smart contract audit,” *Mythx.io*. Accessed on: April 24, 2023. [Online]. Available: https://mythx.io/about/
77. “GPT-4,” *Openai.com*. Accessed on: April 24, 2023. [Online]. Available: https://openai.com/product/gpt-4
78. S. Bubeck et al., “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” *arXiv*, 2023. doi: 10.48550/ARXIV.2303.12712.
79. D. Guido, “Codex (and GPT-4) can’t beat humans on smart contract audits,” *Trail of Bits Blog*. Accessed on: April 24, 2023. [Online]. Available: https://blog.trailofbits.com/2023/03/22/codex-and-gpt4-cant-beat-humans-on-smart-contract-audits/

Received 12.05.2023

INFORMATION ON THE ARTICLE

Yaroslaw Yu. Dorogyy, ORCID: 0000-0003-3848-9852, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine, e-mail: argusyk@gmail.com

Vadym Yu. Kolisnichenko, ORCID: 0009-0009-6472-2807, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine, e-mail: vadym.kolisnichenko@gmail.com

АНАЛІЗ БЛОКЧЕЙН-ТРАНЗАКЦІЙ: КОМПЛЕКСНИЙ ОГЛЯД ЗАСТОСУВАНЬ, ЗАВДАНЬ ТА МЕТОДІВ / Я.Ю. Дорогий, В.Ю. Колісніченко

Анотація. Аналіз блокчейн-транзакцій є потужним інструментом для отримання інформації про дії та поведінку учасників у блокчейн-мережах. Розглянуто застосування, завдання та методи, пов’язані з аналізом блокчейн-транзакцій. Розглянуто різні способи використання аналізу транзакцій, починаючи від його інструментальної ролі в розробленні блокчейн-систем і закінчуючи його ключовим значенням у сфері кримінальних розслідувань. Із використанням загальних методів і технології, що застосовуються у ході такого аналізу, розкрито приховані уявлення та знайдено інформацію, яка є неочевидною. Мета рукопису – всебічний погляд на важливе значення аналізу блокчейн-транзакцій із розкриттям його ключової ролі у криптовалютній індустрії та широкий спектр застосувань поза нею.

Ключові слова: блокчейн-транзакції, аналіз транзакцій, відстеження транзакцій, аналіз потоків, блокчейн криміналістика.

ТЕОРЕТИЧНІ ТА ПРИКЛАДНІ ПРОБЛЕМИ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ ПІДТРИМАННЯ ПРИЙНЯТТЯ РІШЕНЬ

УДК 004.02, 004.67, 004.891.3
DOI: 10.20535/SRIT.2308-8893.2023.4.04

МОДЕЛЮВАННЯ ДИНАМІКИ РИНКУ КРИПТОВАЛЮТ З ВИКОРИСТАННЯМ ІНСТРУМЕНТІВ МАШИННОГО НАВЧАННЯ

Д. МАРТЬЯНОВ, Я. ВИКЛЮК, М. ФЛЕЙЧУК

Анотація. Проаналізовано динаміку кон'юнктури ринку криптовалют (Bitcoin) з використанням інструментарію економетричного оцінювання на основі моделей машинного навчання. Удосконалено метод прогнозування на основі декомпозиції часових рядів та лагових зміщень фінансових індикаторів. Побудовано ансамбль моделей короткочасного прогнозу курсу Bitcoin та проаналізовано його точність порівняно з окремими складовими моделями. Використано моделі часових рядів на основі розрахованих фінансових індикаторів (ADODS, NATR, TRANGE, ATR, OBV, RSI, ADTV). Абсолютне відхилення короткочасного прогнозу склало 9,5$ що становить 0,06% від абсолютного значення.

Ключові слова: ансамблі моделей, машине навчання, часовий ряд, криптовалюта.

АКТУАЛЬНІСТЬ

На сучасному етапі фінансові ринки розвиваються надзвичайно високими темпами, зокрема ринок криптовалют; економетричне моделювання може стати основою для визначення факторів впливу на ціни та обсяги торгів криптовалютами. Таке моделювання передбачає збирання відповідних даних, визначення структури моделі, оцінювання її параметрів і проведення відповідного статистичного аналізу. Економетричні моделі можуть ґрунтуватися на різних методологіях, таких як аналіз часових рядів, регресійний аналіз чи аналіз панельних даних.

Ключові етапи моделювання ринку криптовалют: *Збирання даних.* З цією метою можливе використання ретроспективних даних про ціни на криптовалюту, обсяги торгів, ринкові та фінансові показники та інші змінні. Ці дані можна отримати з різних джерел (біржі криптовалют, постачальників фінансових даних і публічних баз даних навіть з можливістю їх комерційного використання).

Специфікація моделі. Визначення структури моделі, що окреслює зв'язок між залежною змінною (наприклад, ціною криптовалюти) і незалежними змінними (наприклад, ринковими та фінансовими показниками, іншими економічними факторами). Цей етап може передбачати вибір відповідних економетричних методів, таких, як моделі авторегресійної інтегрованої ковзної середньої (ARIMA) [1], моделі векторної авторегресії (VAR) [2] або алгоритми машинного навчання [3].

Оцінювання параметрів. Цей крок передбачає адаптацію моделі до ретроспективних даних та оцінювання відповідних коефіцієнтів, які окреслюють зв'язки між змінними.

Оцінювання моделі. Оцінювання відповідності оцінюваної моделі. Для оцінювання продуктивності моделі можна використовувати різноманітні статистичні тести, такі як перевірка гіпотез, діагностика моделі та перевірка поза вибіркою (наприклад, F-критерій Фішера [4], рівень статистичної похибки [5], T-критерій Ст'юдента [6], Тест Гренджера [7], Тест Дарбіна–Уотсона [8] тощо).

Прогнозування та аналіз. Після успішно застосованих описаних вище етапів модель можна використовувати для прогнозування та аналізу сценаріїв або ж відсіяти.

Указані кроки та побудова дієвої моделі допомагають зрозуміти потенційні майбутні тенденції та поведінку ринку криптовалют на основі оцінених взаємозв'язків.

Важливо зазначити, що економетричні моделі ґрунтуються на історичних даних і припущеннях, а їх точність залежить від обмежень і невизначеностей. Ринок криптовалют відомий своєю нестабільністю та чутливістю до зовнішніх факторів, таких як нормативні зміни, технологічний прогрес, інфляційні або девальваційні очікування (фіктивні змінні), інші тенденції ринку, які можуть створювати проблеми для економіко-математичного моделювання.

Для застосування економетричного моделювання з метою прогнозування тенденцій ринку криптовалют необхідно мати глибинне розуміння специфічних методів статистичного аналізу та особливостей динаміки досліджуваного ринку.

Аналіз останніх досліджень та розробок у даній галузі може допомогти підвищити точність і надійність моделей.

АНАЛІЗ ОСТАННІХ ПУБЛІКАЦІЙ

З активізацією ринку криптовалют значна кількість вітчизняних та зарубіжних учених зацікавилися питаннями аналізу та прогнозування цього ринку з використанням сучасних методів та інструментів. З одного боку, масштабне глобальне зацікавлення цим ринком фінансовими трейдерами, з огляду на можливість одержання значної маржі у короткостроковому періоді та доступ до масштабних баз даних (динамічних датасетів з поділом на численні часові лаги) створює істотні перспективи для економетричного моделювання; на такий тип дослідження існує значний попит в аналітичних колах. Із другого боку, високий рівень волатильності та вплив багатьох суб'єктивних факторів, що важко піддаються екстраполяції, потребує від аналітиків даних використання комплексного підходу та застосування різноманітних методів і інструментів для об'єктивного обґрунтування прикладних моделей прогнозування динаміки ринку криптовалют та миттєвого прийняття рішень трейдерами, що подекуди буває непростим завданням щодо вирішення.

Питанням прогнозування курсу криптовалют, зокрема Bitcoin, з урахуванням низки факторів, що впливають на його вартість, а також окресленню щоденних тенденцій на ринку Bitcoin присвячено працю G. Gurupradeep, M. Harishvaran, K. Amsavalli [9]. У згаданій праці для прогнозування ціни закриття наступного дня, враховано такі фактори, як ціна відкриття, найвища ціна, найнижча ціна, ціна закриття, обсяг Bitcoin, обсяг інших валют і зважена ціна. При цьому автори використовують інструменти Scikit-Learn і моделі «випадкового лісу» і прогнозування.

Л. Кібальник у своїх дослідженнях [10] підтверджує, що ринок криптовалют характеризується значною волатильністю, курсовими коливаннями, складністю застосування адміністративних методів регулювання та кризовими явищами. Тому, як вважають автори, використання традиційних методів моделювання є неефективним, оскільки з викорстанням класичних методів аналізу досить важко отримати адекватні прогнози стосовно розвитку цього ринку. Дослідники пропонують застосування інструментарію фрактального аналізу та аналізу динаміки волатильності, що дозволяє здійснювати постійний моніторинг стану ринку та прогнозування динаміки криптовалют різного ступеня капіталізації.

Цікавим видається науковий підхід L. Catania та S. Grassi [11]. Науковці також відзначають, що дослідження фінансових часових рядів криптовалют досить складно піддається моделюванню, демонструючи екстремальні спостереження, асиметрії та часто нелінійні характеристики, які важко прогнозувати. Автори розробляють динамічну модель нового типу, здатну врахувати довгострокову пам'ять і асиметрію в процесі волатильності, а також наявність змінних у часі асиметрії та ексцесу. Емпіричне дослідження, виконане на великому масиві реальних даних щодо наборів криптовалют, засвідчує докази наявності тривалої пам'яті та ефекту кредитного плеча, що можна вважати вагомим внеском у теорію динаміки волатильності. Такі результати є важливими для управління інвестиційними активами та ризиками, пов'язанами з цим процесом.

Таким чином, аналіз та моделювання ринку криптовалют розкриває важливі наукові горизонти для застосування сучасного інструментарію економетричного прогнозування на фінансових ринках у прикладній площині.

ПОСТАНОВКА ЗАВДАННЯ

З огляду на викладене вище, сформулюємо **мету** дослідження, як побудову ансамблевої моделі короткочасного прогнозу курсу Bitcoin на базі історичних даних та на фінансових індикаторах цієї криптовалюти. У роботі використано історичні дані з розбиттям по одній хвилині за 11,5 днів. Усього набір даних містив 18 056 записів і складався з таких змінних: «price» (курс криптовалюти), «volume» (обсяг трансакції), «count» (кількість трансакцій), «open» (курс на початок торгів), «high» (рівень максимального курсу), «close» (курс криптовалюти на завершення певного періоду), «low» (рівень мінімального курсу). Слід зазначити, що у наявному реальному ряду даних окремі з них були пропущені, проте період відсутності даних не перевищував трьох хвилин. Тому для подальшого аналізу відсутні дані були заповнені з використанням методу інтерполяції з використанням лінійної регресії. Результат інтерполяції наведено на рис. 1.

Як відомо, моделі прогнозування курсу валют можна поділити на дві категорії.

1. *Модель часових рядів*. Досліджується значення криптовалюти з урахуванням її значень за попередні періоди часу без урахування інших факторів [12].

2. *Модель на основі фінансових індикаторів.* Курс криптовалюти прогнозується на основі фінансових індикаторів, які у своїй природі враховують часові затримки та характеристики криптовалюти за попередні моменти часу [13].

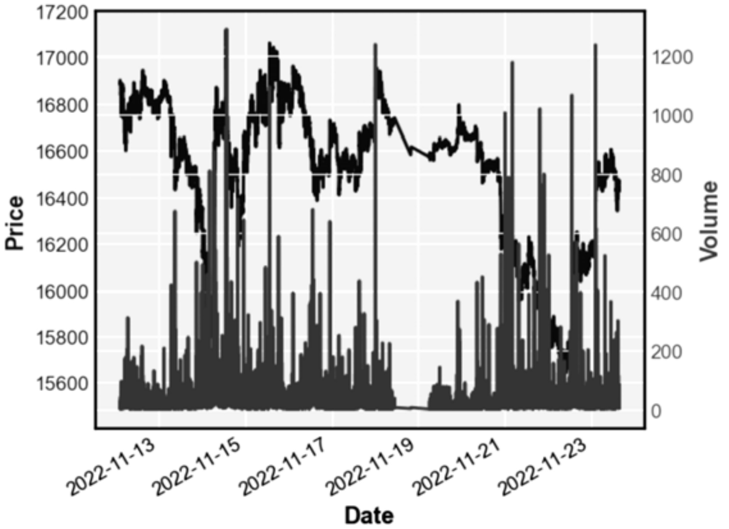


Рис. 1. Динаміка курсу та обсягів продажу Bitcoin, 2022-11-13 – 2022-11-23
Розраховано авторами.

У роботі поєднано ці два підходи та використано різноманітні методи машинного навчання, а також поєднано найбільш дієві з них в ансамбль.

Алгоритм розрахунку:

1) розрахунок фінансових індикаторів;

2) часова декомпозиція цільового поля;

3) визначення лагових затримок та формалізація моделі;

4) побудова та аналіз множини моделей машинного навчання;

5) побудова прогнозу на основі ансамблю моделей.

РОЗРАХУНОК ФІНАНСОВИХ ІНДИКАТОРІВ

Як досліджувані індикатори обрано найбільш популярні та вживані характеристики кон'юнктури ринку криптовалют: ADODS, NATR, TRANGE, ATR, OBV, RSI, ADTV.

Коротко викладемо основні підходи до розрахунку цих фінансових індикаторів.

ADODS — Chaikin A/D Oscillator [14].

Цей індикатор на основі обсягу для вимірювання сукупного грошового потоку. Індикатор припускає, що ступінь тиску купівлі або продажу можна визначити за розташуванням закриття з урахуванням значень максимуму та мінімуму обмінного курсу за досліджуваний період.

Крива ADODS — це загальна сума кожного періоду обсягу грошових потоків (MVF):

$$MFV = \frac{(Close - Low) - (High - Close)}{High - Low} \times \mathrm{V}olume\,;$$

$$ADODS_{P+1} = ADODS_P + MFV\,,$$

де p — індикатор періоду.

ATR Normalized (NATR) [15].

Цей індикатор використовується в технічному аналізі для вимірювання рівня волатильності і визначається за рівняннями:

$$NATR = \frac{100 \times ATR}{Close}\,;$$

$$ATR_{p+1} = \frac{ATR_P\left(n-1\right) + TR}{n}\,, \qquad (1)$$

де p — індикатор періоду; n — кількість періодів;

$$ATR_P = \frac{1}{n}\sum_{i=1}^{n} TR_i\,;$$

$$TR = \max[(High_p - Low_p), \left|High_p - Close_{p-1}\right|, \left|Low_p - Close_{p-1}\right|], \qquad (2)$$

де n — аналізована кількість періодів часу; p — індикатор періоду; TR — діапазон торгівлі.

- *Ttue Range* (TRANCE) [16] — це технічний індикатор, що вимірює денний діапазон курсу плюс будь-який розрив від ціни закриття попереднього дня ((2)).
- *Середній справжній діапазон* (*ATR*) [17] — це індикатор технічного аналізу, який вимірює волатильність ринку шляхом розкладання всього діапазону ціни активу за цей період (рівняння (1)).
- *Балансовий обсяг* (*OBV*) [18] — це технічний індикатор імпульсу торгівлі, який використовує потік обсягу для прогнозування змін курсу акцій:

$$OBV_{p+1} = OBV_p + \begin{cases} Volume: & Close_{p+1} > Close_p; \\ 0: & Close_{p+1} = Close_p; \\ Volume: & Close_{p+1} < Close_p, \end{cases}$$

де p — індикатор періоду.

- *Індекс відносної сили* (*RSI*) — широко використовуваний індикатор осцилятора. Для трейдерів із середньою реверсією він може генерувати сигнали для визначення рівня перекупленої чи перепроданої ціни. RSI також можна використовувати для визначення сили руху тренду:

$$RSI_{step\ one} = 100 - \left[\frac{100}{1 + \frac{Avg\ Gain}{Avg\ Loss}}\right],$$

де $Avg\,Gain$ — середній прибуток (в абсолютному вираженні); $Avg\,Loss$ — середній збиток в абсолютному вираженні.

- *Середній прибуток* ($Avg\,Gain$) *або збиток* ($Avg\,Loss$) [19], що використовується у цьому методі, є середнім відсотковим приростом або збитком відповідно впродовж досліджуваного періоду. У формулі використовується абсолютне значення обсягу середніх втрат. Періоди із втратою ціни враховуються як нуль у розрахунках середнього прибутку. Лаги з тенденцією до підвищення цін враховуються як нуль у процесі розрахунку середнього збитку. Стандартна кількість періодів, які використовуються для розрахунку початкового значення RSI, становить 14.

За умови коли буде розраховано значення 14 періодів, можна переходити до виконання наступного етапу, метою якого є згладжування результатів таким чином, щоб RSI лише наближався до 100 чи до нуля на чітко окресленому трендовому ринку:

$$RSI_{step\,two} = 100 - \left[\frac{100}{1 + \dfrac{(Avg\,Gain)_p \times 13 + (Avg\,Gain)_{p+1}}{(Avg\,Loss)_p \times 13 + (Avg\,Loss)_{p+1}}} \right].$$

- *Середній щоденний обсяг торгів (ADTV)* — це середній обсяг продажу певної валюти впродовж дня. Середній щоденний обсяг торгів є важливим показником, оскільки високий або ж низький обсяг торгів приваблює різні типи трейдерів та інвесторів. Багато трейдерів та інвесторів віддають перевагу вищому рівню середнього щоденного обсягу торгів порівняно з низьким, оскільки за умови великого обсягу легше відкривати та виходити з позицій. Активи з невеликим обсягом мають менше покупців і продавців, і тому, в цьому випадку, може бути важче увійти або вийти на торги на рівні бажаної ціни:

$$ADTV = \frac{Volume_{daily}}{\sum trades}.$$

ЧАСОВА ДЕКОМПОЗИЦІЯ ЦІЛЬОВОГО ПОЛЯ

Як декомпозицію цільового поля у дослідженні обрано адитивну модель (11) [20]:

$$price(t) = Trend\,(t) + Seasonal\,(t) + Residial\,(t)\,.$$

Трендову залежність $Trend\,(t)$ [21] розраховано як лінійну згортку методом найменших квадратів, екстрапольованим на обох кінцях [22].

Сезонна компонента $Seasonal(t)$ визначалась методом ковзного середнього із періодом один день (1440 хв) [23].

Результат до композиції подано на рис. 2. Як видно з рисунка, на графіку взаємозв’язку курсу Bitcoin з досліджуваними фінансовими індикаторами спостерігаються чіткі трендові та сезонні компоненти залежності з невеликими розривами у період вихідних днів. У результаті декомпозиції компонента $\mathrm{Re}\,sidial(t)$ виступатиме в ролі цільового поля.

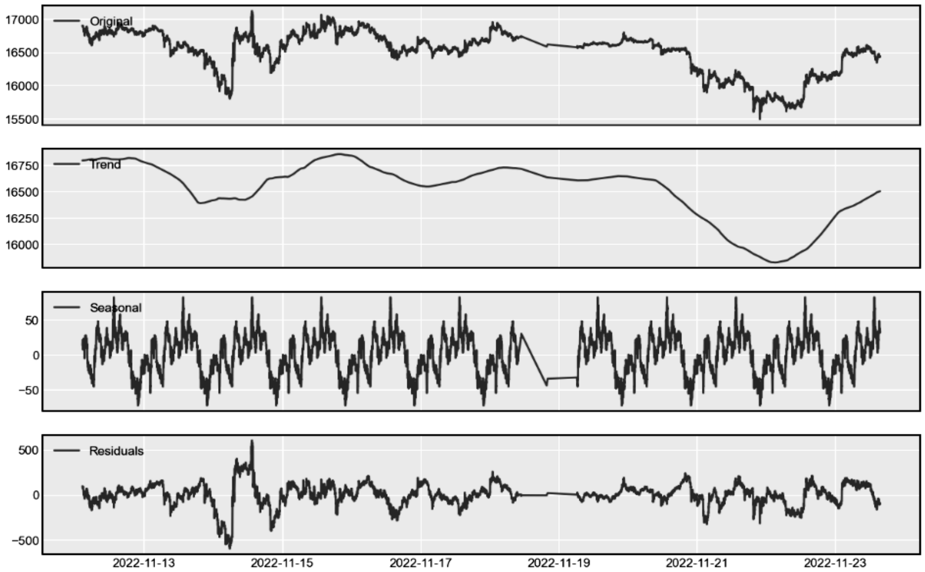


Рис. 2. Декомпозиція часового ряду курсу Bitcoin, 2022-11-13 – 2022-11-23. *Розраховано авторами.*

Порівняння динаміки фінансових індикаторів з отриманою випадковою компонентою показано на рис. 3.

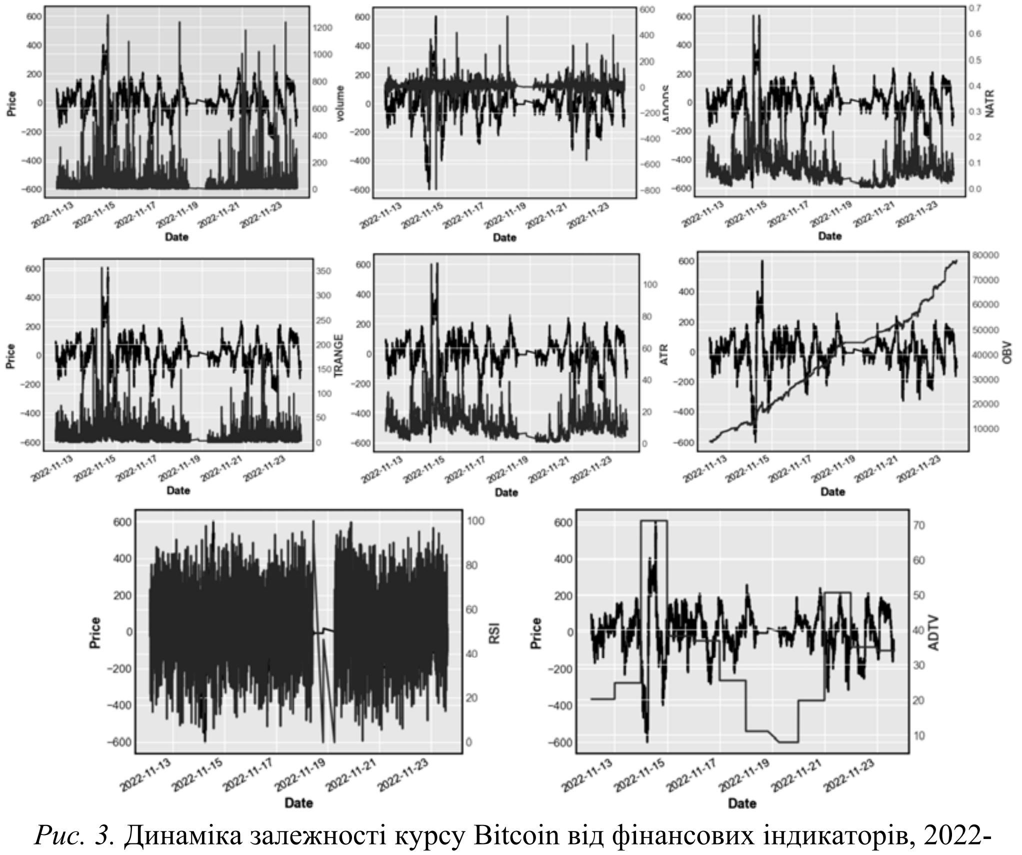


Рис. 3. Динаміка залежності курсу Bitcoin від фінансових індикаторів, 2022-11-13 – 2022-11-23
Розраховано авторами

Як видно з рис. 3, динаміка досліджуваної залежної змінної від фінансових індикаторів суттєво розрізняються і потребують подальшого аналізу та встановлення наявності природи і типу функціональної залежності.

ВИЗНАЧЕННЯ ЛАГОВИХ ЗАТРИМОК ТА ФОРМАЛІЗАЦІЯ МОДЕЛІ

Додатково, для аналізу наявності факту і напряму зв'язку між фінансовими індикаторами та цільовим полем, виконано кореляційний аналіз (рис. 4).

Як видно з аналізу, між полями ATR та NATR існує сильний лінійний зв'язок, тому для подальшого аналізу залишено поле NATR, оскільки воно являє собою нормалізоване ATR. Також існує зв'язок між полем Volume та TRANGE. Цей зв'язок є на межі лінійного, тому прийнято рішення залишити ці два поля для подальшого детальнішого аналізу. Окрім того, можна побачити, що випадкова компонента майже не корелює лінійно з жодним із зазначених фінансових індикаторів. Це свідчить або про відсутність лінійного зв'язку та необхідність використання нелінійних моделей, або про необхідність побудови комплексної лінійної моделі, що включає в себе всі зазначені фінансові індикатори (багатофакторної регресійної моделі). Проте, як показує практика трейдингу на ринку криптовалют, саме визначені нами індикатори найчастіше враховуються трейдерами для прийняття швидких біржових рішень.

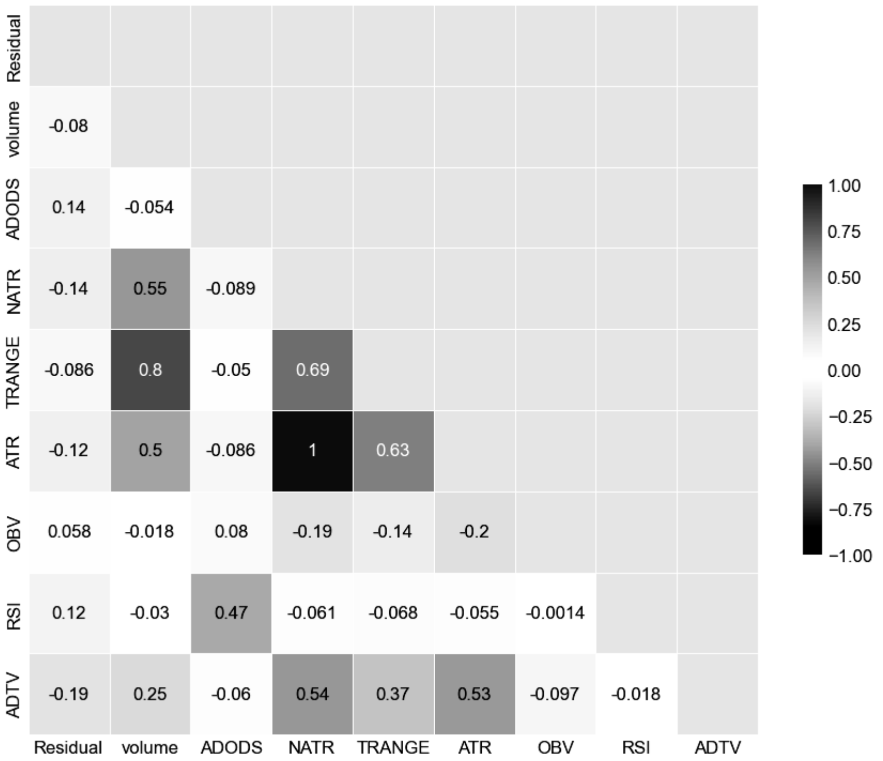


Рис. 4. Матриця кореляційного зв'язку між фінансовими індикаторами динаміки ринку криптовалют та курсом Bitcoin, 2022-11-13 – 2022-11-23
Розраховано авторами

Для врахування впливу лагової затримки випадкової компоненти $(Resudial(t))$ виконано автокореляційний аналіз у межах сезонної компоненти, тобто на 1440 хв. Результат аналізу подано на рис. 5.

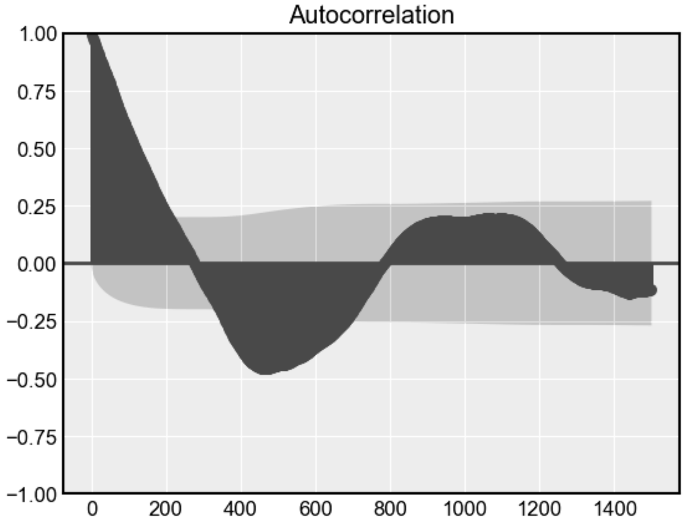


Рис. 5. Автокореляційний аналіз для випадкової компоненти $(Resudial(t))$, 2022-11-13 – 2022-11-23
Розраховано авторами

Як видно з рисунка, лагова кореляція істотно знижується до нуля і обернено зростає в межах лагу 465 хв (7 год 45 хв), що відповідає сценарію класичних торгів на ринку криптовалют. Для уточнення необхідного лагу та усунення взаємної кореляції виконано частковий автокореляційний аналіз. Результат ілюструє рис. 6.

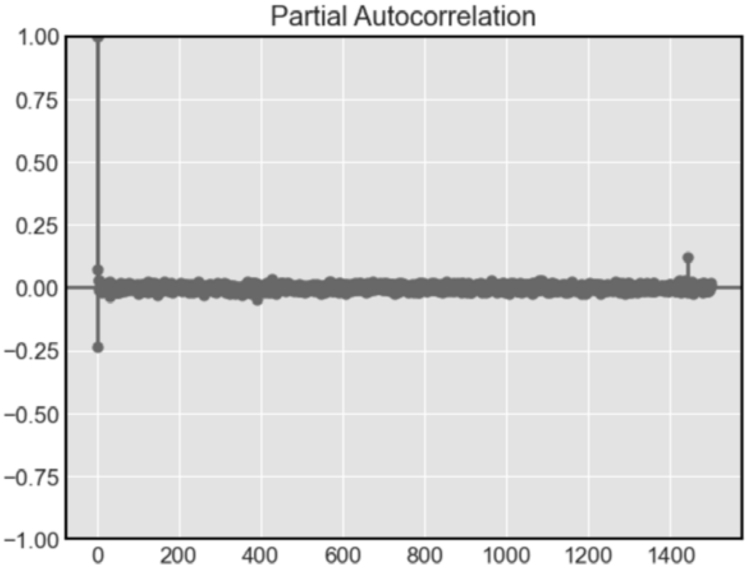


Рис. 6. Частковий автокореляційний аналіз для випадкової компоненти $(Resudial(t))$, 2022-11-13 – 2022-11-23
Розраховано авторами.

Як видно з рис. 6, найвищого рівня кореляційний зв'язок притаманний для компоненти з лаговою затримкою 1, 2 та 1440. Для встановлення наявності лагових затримок між іншими фінансовими індикаторами та випадковою компонентною також виконано відповідний автокореляційний аналіз (рис. 7). Як видно з рисунка, всі коефіцієнти кореляції є дуже низькими, тому немає сенсу враховувати лагову затримку під час побудови моделей.

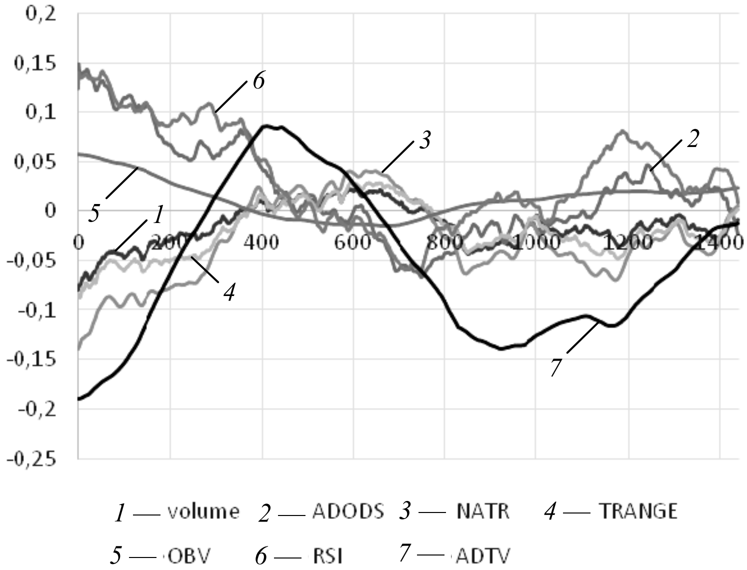


Рис. 7. Залежність коефіцієнта кореляції між випадковою компонентою ($Resudial(t)$) та фінансовими індикаторами з відповідним часовим лагом, 2022-11-13 – 2022-11-23
Розраховано авторами

Це пояснюється специфікою викоаного аналізу. Оскільки йдеться про розбиття даних по одній хвилині, то логічним є відсутність такого впливу, адже кон'юнктура ринку криптовалют у таких часових лагах дуже швидко змінюється і часто трейдери враховують більші часові лаги.

Таким чином, у результаті проведеного аналізу формалізовану модель прогнозу курсу Bitcoin можна подати у такому вигляді:

$$BIT_{USD}(t+1) = Trend(t) + Seasonal(t) + F(Resudial(t), Resudial\ (t-1),$$
$$Resudial(t-2), Resudial(t-465), Resudial(t-1440), volume(t), ADODS(t),$$
$$NATR(t), TRANGE(t), OBV(t), RSI(t), ADTV(t)\ . \qquad (3)$$

Як видно з рівняння (3), курс Bitcoin можна розраховувати як суму трендової, сезонної та випадкової компонент, яка прогнозується за допомогою ансамблю моделей.

ПОБУДОВА ТА АНАЛІЗ МНОЖИНИ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Одним із завдань дослідження є побудова моделі для прогнозу курсу валют на майбутні періоди. Модель, подана у вигляді рівняння (3), дає змогу прогнозувати лише на один період наперед. Для побудови прогнозу на декілька майбутніх періодів можна застосувати два підходи:

1. Для кожного із факторів побудувати прогнозну модель на один період наперед і відповідно використовувати прогнозні значення як вхідні для прогнозування на наступний період часу (ланцюговий метод).

Як вихідне поле використовувати значення $Residial(t+\text{lag})$ зміщеного на необхідні періоди наперед.

Перший підхід потребує побудови великої кількості моделей і зумовить накопичення помилки за ітерацій на кожний наступний період. Другий підхід дозволить зменшити накопичувальну помилку, але потребує побудови окремої моделі для кожного лагу прогнозування. Тому для розрахунків обрано саме другий підхід.

Як тестові моделі виокремлено: лінійну множинну регресію [24] (Linear), регресію Губара [25] (Huber) та нейронну мережу зворотного поширення помилки [26] (MLP), що складається із двох прошарків по 100 нейронів кожен. Для тестування точності та адекватності моделей використано кросвалідацію з величиною розбиття 3. Результати оцінювання точності прогнозування наведено в табл. 1.

Як видно з таблиці, точність прогнозу у випадку тестових даних є досить високою. Як і слід було очікувати, зі збільшенням часу прогнозу точність моделей поступово зменшується. Однак навіть на 10 періодів наперед вона є досить висока.

Таблиця 1. Коефіцієнти кореляції моделі прогнозу (R^2) для різних лагів для $Resudial(t+\text{lag})$, 2022-11-13 – 2022-11-23

Lag	Linear	MLP	Huber	VotingRegressor
0	0,99	0,99	0,99	0,99
1	0,98	0,98	0,98	0,98
2	0,97	0,96	0,97	0,97
3	0,96	0,96	0,96	0,96
4	0,96	0,94	0,95	0,95
5	0,95	0,92	0,95	0,95
6	0,94	0,93	0,94	0,94
7	0,93	0,92	0,93	0,93
8	0,92	0,91	0,92	0,92
9	0,91	0,90	0,91	0,91
10	0,90	0,89	0,90	0,90

Розраховано авторами.

Для уникнення випадкових флуктуацій моделей їх об’єднано в ансамбль за допомогою VotingRegressor, який усереднює результати окремих моделей.

ПОБУДОВА ПРОГНОЗУ НА ОСНОВІ АНСАМБЛЮ МОДЕЛЕЙ

Для побудови прогнозу набір даних було поділено на навчальний та тестовий в пропорції 90/10. Як тестовий набір обрано дані за останній період. Результати навчання і тестування наведено в табл. 2.

Таблиця 2. Точність навчання (R^2) на навчальному та тестового набор даних для $Residial(t + \mathrm{lag})$

Lag	Linear		MLP		Huber		VotingRegressor	
	Train	Test	Train	Test	Train	Test	Train	Test
0	0,99	0,99	0,99	0,95	0,99	0,99	0,99	0,99
1	0,99	0,98	0,99	0,98	0,98	0,98	0,99	0,98
2	0,98	0,97	0,98	0,97	0,97	0,97	0,98	0,97
3	0,97	0,96	0,97	0,94	0,97	0,96	0,97	0,96
4	0,96	0,95	0,94	0,85	0,96	0,95	0,96	0,95
5	0,95	0,94	0,93	0,88	0,95	0,94	0,95	0,90
6	0,94	0,93	0,94	0,89	0,94	0,93	0,94	0,92
7	0,93	0,92	0,93	0,92	0,93	0,92	0,93	0,92
8	0,92	0,91	0,92	0,89	0,92	0,91	0,92	0,91
9	0,91	0,90	0,91	0,89	0,91	0,90	0,91	0,89
10	0,90	0,89	0,91	0,86	0,90	0,89	0,90	0,87

Розраховано авторами.

Як видно з таблиці, точність навчання поступово спадає з лагом прогнозу. Чим менша затримка в прогнозі, тим точніший результат. У розглядуваному випадку, кожен лаг становить 1 хв. Отже, чим швидше буде прийнято рішення, тим менша помилка в його прийнятті. Це точність прогнозу для випадкової компоненти числового ряду. Для визначення точності моделі (3) необхідно до цих результатів додати сезонну та трендову компоненти і побудувати прогноз реального курсу валют. Прогноз курсу Bitcoin для тестових значень подано на рис. 8.

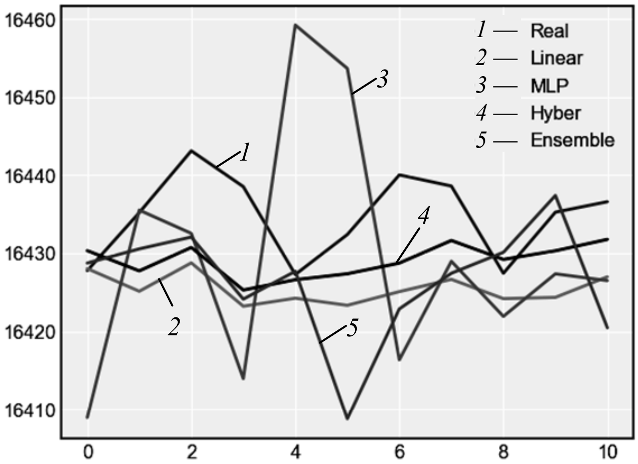


Рис. 8. Прогноз курсу Bitcoin згідно з розробленим ансамблем моделей на 10 хв наперед порівняно з тестовими даними
Розраховано авторами

Як видно з рисунка, значення мають досить сильне відхилення від реальних даних, однак розрахунок середнього абсолютного відхилення становить 9,5, що становить 0,06% від абсолютного значення. Це може свідчити про достатній рівень точності моделі.

ВИСНОВКИ

1. На сучасному етапі фінансові ринки розвиваються надзвичайно високими темпами, зокрема і ринок криптовалют, для якого економетричне моделювання може стати основою для визначення факторів впливу на його кон'юнктуру. Попри істотні можливості застосування економетричного інструментарію для аналізу цього ринку (генерування об'ємних даних у режимі реального часу), слід вказати й на певні обмеження такого аналізу, зокрема істотна волатильність ринку та часто суб'єктивний чи інтуїтивний характер прийняття рішень трейдерами криптовалют.

2. Для оперативного прийняття рішень щодо купівлі-продажу валютних активів на фондових біржах запропоновано застосування ансамблевої моделі короткочасного прогнозу курсу валюти, ґрунтуючись на історичних даних основних характеристик ринкової кон'юнктури (курс криптовалюти, обсяг транзакції, кількість транзакцій, курс на початок торгів, рівень максимального курсу, курс криптовалюти на завершення певного періоду, рівень мінімального курсу та фінансових індикаторах цієї криптовалюти (ADODS, NATR, TRANGE, ATR, OBV, RSI, ADTV).

3. У процесі аналізу моделі прогнозування курсу криптовалют можна використати дві категорії моделей: модель часових рядів; модель на основі фінансових індикаторів.

4. Визначено ключові етапи моделювання ринку криптовалют: 1) збирання даних; 2) специфікація моделі; 3) оцінювання параметрів; 4) оцінювання моделі; 5) прогнозування та аналіз.

5. Серед основних кроків для побудови моделі для аналізу ринку криптовалют рекомендовано застосовувати: часову декомпозицію цільового поля; визначення лагових затримок та формалізацію моделі; побудову та аналіз множини моделей машинного навчання; побудову прогнозу на основі ансамблю моделей.

6. Основні висновки з виконаного аналізу:

- у взаємозв'язку курсу Bitcoin з досліджуваними фінансовими індикаторами спостерігаються чіткі трендові та сезонні компонентні залежності з невеликими розривами у період вихідних днів;

- найміцніший зв'язок визначено між такими змінними, як ATR і NATR, а також Volume та TRANGE. Цей зв'язок є на межі лінійного, тому прийнято рішення залишити ці два поля для подальшого детальнішого аналізу. Відзначено, що випадкова компонента майже не корелює лінійно з жодним із зазначених фінансових індикаторів. Це свідчить або про відсутність лінійного зв'язку та необхідність використання нелінійних моделей, або про потребу побудови комплексної лінійної моделі, що включає в себе основні фінансові індикатори (багатофакторної регресійної моделі);

- установлено, що лагова кореляція істотно знижується до нуля і обернено зростає в межах лагу 465 хв (7 год 45 хв), що відповідає сценарію класичних торгів на ринку криптовалют; найвищий рівень кореляційного зв'язку притаманний для компоненти з лаговою затримкою 1, 2 та 1440; оскільки всі коефіцієнти кореляції між випадковою компонентою та фінансовими індикаторами є надто низькими, тому немає сенсу враховувати лагову затримку у побудові такого типу моделей; курс Bitcoin можна розраховувати як суму трендової, сезонної та випадкової компонент, яка прогнозується за допомогою ансамблю моделей.

ЛІТЕРАТУРА

1. Gandhi Pratik, *7 Statistical Tests to validate and help to fit ARIMA model*. Available: https://towardsdatascience.com/7-statistical-tests-to-validate-and-help-to-fit-arima-model-33c5853e2e93
2. Kotzé Kevin, *Vector autoregression models*. Available: https://kevinkotze.github.io/ts-7-var/.
3. Taiwo Oladipupo Ayodele, *Types of Machine Learning P.19-28. Algorithms*. Available: https://cdn.intechopen.com/pdfs/10694/InTech-Types_of_machine_learning_algorithms.pdf
4. Onchiri Sureiman and Callen Moraa Mangera, "F-test of overall significance in regression analysis simplified," *Journal of the Practice of Cardiovascular Sciences*, vol. 6, issue 2, pp. 116–122, May-August 2020. doi: 10.4103/jpcs.jpcs_18_20.
5. Sander Greenland et al., "Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations," *The American Statistician, Online Supplement*, pp. 1–12, 2016. Available: https://events.gwdg.de/event/482/attachments/391/580/anwer_to_the_ASA_statement_misinterpretations.pdf
6. L. Brown, "The conditional level of Student's – test," *The Annals of Mathematical Statistics*, vol. 38, no. 4, pp. 1068–1071, Aug., 1967. Available: https://faculty.wharton.upenn.edu/wp-content/uploads/2012/04/Conditional-level-of-students-t-test.pdf
7. Xiaojun Song and Abderrahim Taamouti, "A better understanding of Granger causality analysis: A big data environment," *Oxford Bulletin of Economics & Statistics*, pp. 2–25, August 2019. Available: https://www.researchgate.net/publication/329803300_A_Better_Understanding_of_Granger_Causality_Analysis_A_Big_Data_Environment
8. Champion Robert and Mills M. Terence, "Demonstrating the Durbin-Watson Statistic," *Journal of the Royal Statistical Society Series D (The Statistician)*, 47(4), pp. 643–644, December 1998. doi: 10.1111/1467-9884.00161.
9. G. Gurupradeep, M. Harishvaran, and K. Amsavalli, "Cryptocurrency Price Prediction using Machine Learning, 2023," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 12, issue 4, pp. 808–814, April 2023. doi: 10.17148/IJARCCE.2023.124140.
10. L.O. Kibalnyk, O.A. Kovtun, and G.B. Danylchuk, "Modeling and analysis of the current state of the cryptocurrency market," *Economics and management in the period of digital transformation of business, society and the state: materials of the Jubilee International Scientific and Practical Conference (May 28–29, 2020, Zaporizhzhia)*. Zaporizhzhia: ZNU Engineering Institute, 2020, pp. 112–117.
11. Leopoldo Catania and Stefano Grassi, "Modelling Crypto–Currencies Financial Time–Series," *SSRN Electronic Journal*, pp. 2–38, August 2017. doi: 10.2139/ssrn.3028486.
12. George E.P. Box, *Time series analysis: forecasting and control*; fifth edition, 2016, 668 p. Available: http://repo.darmajaya.ac.id/4781/1/Time%20Series%20 Analysis_%20Forecasting%20and%20Control%20%28%20PDFDrive%20%29.pdf
13. Andrea Majdáková, Blanka Giertliová, and Iveta Hajdúchová, "Prediction by financial and economic analysis in the conditions of forest enterprises," *Journal of Forest Science*, 66, 2020 (1), pp. 1–8. Available: https://www.agriculturejournals.cz/pdfs/jfs/2020/01/01.pdf
14. *Chaikin A/D Oscillator*. Available: https://www.cryptohopper.com/resources/ technical-indicators/283-chaikin-a-d-oscillator.
15. *Normalized Average True Range*. Available: https://taapi.io/indicators/normalized-average-true-range/
16. *Average True Range*. Available: https://www.fidelity.com/learning-center/trading-investing/technical-analysis/technical-indicator-guide/atr.

17. *Average True Range*. Available: https://www.wallstreetmojo.com/average-true-range/
18. William Wai Him Tsang and Terence Tai Leung Chong, “Profitability of the On-Balance Volume Indicator,” *Economics Bulletin*, 29(3), pp. 2424–2431, January 2009.
19. *Average Rate Of Return: Meaning, How To Calculate And Uses*. Available: https://in.indeed.com/career-advice/career-development/average-rate-of-retur
20. Kramar Vadim and Alchakov Vasiliy, “Time-Series Forecasting of Seasonal Data Using Machine Learning Methods,” *Algorithms*, vol. 16, issue 5, pp. 2–16, 2023. doi: 10.3390/a16050248.
21. Andrius Buteikis, *Time series with trend and seasonality components*. 66 p. Available: http://web.vu.lt/mif/a.buteikis/wp-content/uploads/2019/02/Lecture_03.pdf
22. Harding Ben and Clayton V. Deutsch, *Trend Modeling and Modeling with a Trend*. May 12, 2021, 14 p. Available: https://geostatisticslessons.com/lessons/trendmodeling
23. R. Shumway and D. Stoffer, *Time series analyses and its applications*; 3rd ed. 2011, 576 p. Available: http://pzs.dstu.dp.ua/DataMining/times/bibl/TimeSeries.pdf
24. Hariaji Joko, *Simple Linear Regression (SLR) Model and Multiple Linear Regression (MLR) Model*, May 2021, pp. 1–34. doi:10.13140/RG.2.2.17237.35044.
25. Qiang Sun and Wenxin Zhou, “Adaptive Huber Regression,” *Journal of the American Statistical Association*. October 2018, 56 p. doi: 10.1080/01621459. 2018.154312. https://www.researchgate.net/publication/317732614_Adaptive_Huber_Regression
26. Dastres Roza and Soori Mohsen, “Artificial Neural Network Systems,” *International Journal of Imaging and Robotics*, 21(2), pp. 13–25, March 2021. Available: https: //www.researchgate.net/publication/350486076_Artificial_Neural_Network_Systems

Надійшла 30.06.2023

INFORMATION ON THE ARTICLE

Dmytro I. Martjanov, ORCID: 0009-0003-3919-4412, Lviv Polytechnic National University, Ukraine, e-mail: d.martjnoff@gmail.com

Yaroslav I. Vyklyuk, ORCID: 0000-0003-4766-4659, Lviv Polytechnic National University, Ukraine, e-mail: vyklyuk@ukr.net

Mariya I. Fleychuk, ORCID: 0000-0002-0417-9066, Stepan Gzhytskyi National University of Veterinary Medicine and Biotechnologies, Ukraine, e-mail: fleychukm@gmail.com

MODELING CRYPTOCURRENCY MARKET DYNAMICS USING MACHINE LEARNING TOOLS / D.I. Martjanov, Ya.I. Vyklyuk, M.I. Fleychuk

Abstract. The article analyzes the dynamics of the cryptocurrency market (Bitcoin) using econometric estimation tools based on machine learning models. The forecasting method is improved based on time series decomposition and lagged shifts of financial indicators. An ensemble of short-term forecast models for the Bitcoin exchange rate is built, and its accuracy is analyzed and compared to individual component models. Time series models are used along with calculated financial indicators (ADODS, NATR, TRANGE, ATR, OBV, RSI, ADTV). The absolute deviation of the short-term forecast amounted to $9.5, which is 0.06% of the absolute value.

Keywords: ensemble models, machine learning, time series, cryptocurrency.

UDC 004.424.4
DOI: 10.20535/SRIT.2308-8893.2023.4.05

SEMI-SUPERVISED INVERTED FILE INDEX APPROACH FOR APPROXIMATE NEAREST NEIGHBOR SEARCH

A. BAZDYREV

Abstract. This paper introduces a novel modification to the Inverted File (IVF) index approach for approximate nearest neighbor search, incorporating supervised learning techniques to enhance the efficacy of intermediate clustering and achieve more balanced cluster sizes. The proposed method involves creating clusters using a neural network by solving a task to classify query vectors into the same bucket as their corresponding nearest neighbor vectors in the original dataset. When combined with minimizing the standard deviation of the bucket sizes, the indexing process becomes more efficient and accurate during the approximate nearest neighbor search. Through empirical evaluation on a test dataset, we demonstrate that the proposed semi-supervised IVF index approach outperforms the industry-standard IVF implementation with fixed parameters, including the total number of clusters and the number of clusters allocated to queries. This novel approach has promising implications for enhancing nearest-neighbor search efficiency in high-dimensional datasets across various applications, including information retrieval, natural language search, recommendation systems, etc.

Keywords: approximate nearest neighbor search, inverted file index, high-dimensional data, machine learning.

INTRODUCTION

Approximate Nearest Neighbor (ANN) [1] search is a fundamental problem in many data-driven applications, spanning domains such as information retrieval, image processing, natural language search, and recommendation systems. The efficient retrieval of similar data points from vast datasets is critical for tasks that involve high-dimensional data representations, where exhaustive search methods become computationally infeasible. As the dataset size grows, the computational cost of performing an exact nearest neighbor search using brute force algorithms becomes prohibitive. Brute force approaches involve comparing each query vector with every data point in the dataset, leading to computational inefficiencies and impractical execution times for large datasets. Approximate nearest neighbor algorithms offer a trade-off between search accuracy and efficiency, allowing for the retrieval of reasonably accurate results within a significantly reduced search space. By intelligently approximating the nearest neighbors, these algorithms enable faster exploration of large datasets, making them essential for real-world applications where timely responses are crucial, such as image and text search, recommendation systems, and similarity-based clustering.

One popular approach in ANN is the Inverted File (IVF) index method [2]. Originally, the IVF index was an inverted indexing technique that partitions the dataset into a set of Voronoi cells or “buckets” [3]. Each bucket corresponds to a cluster of data points, and the indices of data points within each bucket are stored efficiently. During the search process, queries are mapped to their corresponding buckets, and the search is constrained to the nearest neighbors within these buckets, significantly reducing the search space and accelerating the process.

The standard IVF index has shown remarkable performance gains in nearest neighbor search tasks. However, it faces challenges in scenarios with unevenly distributed data, leading to imbalanced bucket sizes [4]. These imbalances can result in a suboptimal trade-off between search efficiency and accuracy, as some buckets might be excessively populated, while others remain underutilized. In addition to challenges posed by unevenly distributed data and imbalanced bucket sizes, another significant issue that the standard IVF index may encounter relates to the formation of centroid clusters. The standard approach typically relies on unsupervised clustering techniques to create the centroids or representatives for each bucket. This process can potentially lead to suboptimal cluster assignments, especially when the training data for centroid formation is insufficient or poorly representative of the underlying data distribution.

To address this limitation, we propose a novel modification to the IVF index method that leverages supervised learning techniques. Specifically, we train classification neural networks to assign query vectors to their most appropriate bucket, based on the similarity to vectors in the dataset. Moreover, we incorporate an optimization objective to minimize the standard deviation of the bucket sizes, further refining the indexing process. By doing so, we aim to achieve more balanced cluster sizes, effectively mitigating the impact of unevenly distributed data.

PRELIMINARIES

Let's formulate a general ANN problem. Let $X=\{x_i \epsilon \mathbb{R}^d \mid i=\overline{1,N}\}$ be a set of N d-dimensional vectors representing the data points in the dataset. The objective of ANN search is to efficiently find, for a given query vector $q \epsilon \mathbb{R}^d$, an approximate nearest neighbor $x^* \epsilon X$ such that the distance between q and x^* is minimized.

In the Inverted File Index (IVF) approach, we partition the dataset X into K disjoint subsets or buckets, denoted as $B_1, B_2 \dots B_K$. Each bucket corresponds to a subset (cluster) of vectors in X with corresponding centroids c_i — centroid of corresponding B_i.

The ANN search with the IVF index can be formulated as follows. Given the metric function *dist*, a query vector $q \in \mathbb{R}^d$, the goal is to find the bucket B_{query}, with a corresponding centroid c_{query} that minimizes the distance to the query vector — equation:

$$c_{query} = \underset{\{c_1..c_k\}}{\text{argmin}} \ (dist\ (q, c_i)) \ .$$

Once the bucket B_{query}, is identified, we need to find x^* — approximate nearest neighbor within that bucket using brute force search — equation:

$$x^* = \underset{x \in B_{query}}{\text{argmin}}(dist(q,x)) \, .$$

Optionally, to improve accuracy, it is possible to use several B_j adjoining to B_{query} buckets on the last step depending on the method hyperparameter set.

SEMI-SUPERVISED INVERTED FILE INDEX APPROACH

Let *dist* — some metric function (euclidian, manhattan, etc.).

Let $X = \{x_i \epsilon \mathbb{R}^d \mid i = \overline{1,N}\}$ vectors representing the data points in the dataset.

Let $Q = \{q_i \epsilon \mathbb{R}^d \mid i = \overline{1,M}\}$ — a set of M d-dimensional vectors with a similar distribution to real-life production queries be a queries training set, $M << N$.

Let $R = \{r_i \epsilon X \mid r_i = \underset{x \in X}{\text{argmin}}(dist(q_i, x)),\ i = \overline{1,M}\}$ — set of ground truth nearest neighbors (responses) from X for each $q_i \in Q$.

Let $K \epsilon \mathbb{N}$ — method hyperparameter, a desired amount of buckets $B_1, B_2, \dots, B_K$, such that $X = \bigcup_{i=1}^{K} B_i$ and $B_i \cap B_j = \varnothing$ if $i \neq j$.

Let $NN : \mathbb{R}^d \to \mathbb{R}^K$ – some vector function — equation:

$$NN_j(q_i) = P\{q_i \epsilon B_j / r_i \epsilon B_j\} \text{ for } j = \overline{1,K}, \qquad (1)$$

where $P\{q_i \in B_j / r_i \in B_j\}$ — is a conditional probability that $q_i \in B_j$ given $r_i \in B_j$. In our case a multi-layer perceptron [5] with a final softmax layer — equation $\text{softmax}_i(z) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$ for $i = \overline{1,K}$, that distributes query vectors q_i into buckets $B_1, B_2, \dots, B_K$. We also want this function to have a specific property, that it distributes query vectors $q_i \in Q$ to the same bucket as their corresponding responses $r_i \in R$.

We can estimate the NN's parameters using the maximum likelihood estimation method [6; 7], if we consider the task as a standard softmax multiclass classification with a cross-entropy loss function — equation $CE(y, \hat{y}) = -\sum_{i=1}^{K} y_i \log(\tilde{y}_i)$. If we consider Q as an input training set and on each epoch step we can calculate actual training targets Y as follows $Y = \{\underset{j=1,K}{\arg\max}(\{NN_j(r_i)\}),\ i = \overline{1,M}\}$ — for each training query we assign its ground truth nearest neighbor's bucket as a target bucket. As a result of NN training, we can explicitly distribute input queries by buckets — equation $bucket(q) = \underset{\boldsymbol{i}=1,\boldsymbol{K}}{\text{argmax}}(\{NN_i(q)\})$ for $q \in \mathbb{R}^d$ and implicitly get the desired buckets $B_1, B_2, \dots, B_K$ — equation:

$$B_j = \left\{ x \epsilon X \mid \underset{i=1,K}{\arg\max}(\{NN_i(x)\}) = j \right\} \text{ for } j = \overline{1,K}. \qquad (2)$$

STANDARD DEVIATION-BASED BUCKET SIZE REGULARIZATION

The vanilla approach proposed in the previous paragraph can produce imbalanced buckets $B_1, B_2, \dots, B_K$ in the result, for example, *NN* will distribute all the query items in the single bucket, so there will be no full power use of the IVF index. If we want the most efficient computational power of the IVF index method, then we obviously need buckets of the most equal size so that the expectation of the search time of a brute force search over a random bucket takes the minimum time. Let $S=\{s_i=|B_i| \mid i=\overline{1,K}\}$ — set of buckets sizes after we have trained NN that distributes query vectors by buckets. We can calculate the standard deviation of the dataset *S*: $\sigma(S)=\sqrt{\left(\frac{\sum(s_i-\bar{s})^2}{N-1}\right)}$. If we want to have buckets of approximately equal sizes then we need to minimize $\sigma(S)$. The problem here is that this function is not differentiable with respect to the parameters of the *NN* model, so we need to use a differentiable approximation of $\sigma(S)$.

Using equations (1), (2) we can calculate the expectation of size for each bucket as follows — equation:

$$\widetilde{s_j}=\sum_{i=1}^{N} NN_j(x_i) \text{ for } x_i\in X; \text{ for } j=\overline{1,K}. \tag{3}$$

So, we can have $\widetilde{S}=\{s_i \mid j=\overline{1,K}\}$ — set of expectations of bucket sizes after we have trained *NN* that distributes query vectors by buckets. And $\sigma(\widetilde{S})$ which is differentiable with respect to the parameters of the *NN* model.

Finally, we can introduce a combined multiclass cross-entropy loss function with std-based bucket size regularization in equation:

$$L(y,\hat{y},X)=\left(-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(\widetilde{y}_{ij})\right)+\gamma*\sigma(\widetilde{S}), \tag{4}$$

where $\left(-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(\widetilde{y}_{ij})\right)$ is a standard cross-entropy component; $\sigma(\widetilde{S})$ — approximated standard deviation of bucket sizes and $\gamma\in[0,+\infty)$ — regularization scale.

TRAINING ALGORITHM

1. Defining K — desired number of buckets and M — desired maximum bucket size.
2. Initialization of multiclass classification *NN* weights [8].
3. On each training epoch:
    1. Calculate current epoch targets $Y=\{\underset{j=1,K}{\operatorname{argmax}}(\{NN_j(r_i)\})\}$.
    2. Calculate the multiclass cross-entropy loss component using $q_i\in Q$ as inputs and $y_i\in Y$ as targets.
    3. Calculate expectations of sizes for each cluster — equation (3).
    4. Calculate $\sigma(\widetilde{S})$ — std-regularization component.
    5. Calculate aggregated loss equation (4).
    6. Do the backpropagation step using stochastic gradient descent modification, for example, Adam [9], and update NN's weights.

4. After the training process is complete, we select the best checkpoint based on the desired performance metric, for example, precision where the actual maximum bucket size < *M*. If there is no such checkpoint in which the maximum actual bucket size is lower than the desired one, then select the checkpoint with the size closest to the desired one and display the corresponding warning.

It could also be useful to apply some dynamic scaling of γ regularization parameter to achieve better precision performance results.

EXPERIMENTAL RESULTS

We've used 3 different configurations in our experiments:

1. Both indexed and query data have a Normal distribution: $X \sim N(0,1)$; $Q \sim N(0,1)$.

2. Both indexed and query data have a skewed Exponential distribution: $X \sim Exponential(1)$; $Q \sim Exponential(1)$.

3. Indexed data has a Normal distribution and query data has an Exponential distribution that can be similar to different life scenarios: $X \sim N(0,1)$; $Q \sim Exponential(1)$.

In all cases we use 64-dimensional vectors. We also split query data Q to training and testing parts equally in order to minimize the risk of overfitting and getting incorrect results — we use the train part during *NN's* weights optimization and the test part to calculate final metrics. We use a three-layer perceptron with tanh activation functions and Adam [9] optimization algorithm using pytorch framework [10]. We evaluate our algorithm compared to a faiss IVF implementation [11] which is a current industrial standard using *SMAPE* and precision metrics — equations:

$$SMAPE(A,F) = 100*\frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{|A_i + F_i|/2};$$

$$Precision = \frac{TP}{TP+FP}.$$

Where in our case A_i is the distance between *i*-th query vector q_i and its actual nearest neighbor from *X* and F_i is the distance between *i*-th query vector q_i and its suggested by algorithm approximate nearest neighbor from *X*. In other words, the SMAPE metric shows us how much the distances to the ground truth nearest neighbors and to the approximated neighbors differ on average.

In the case of the precision metric, we have TP — the number of cases where the approximate nearest neighbor equals the actual nearest neighbor and FP — the number of cases where the approximate nearest neighbor differs from the actual nearest neighbor. In other words, this metric shows us how often our approximated nearest neighbors exactly coincide with the ground truth ones.

We have final results presented in Tables 1, 2, 3. We also have a general structure of the result table:

– *X*-size — number of vectors in the indexed dataset;

– *Q*-size — number of vectors in the queries training set;

– *K* — number of buckets in the algorithm;

– *N*probe — number of adjoining buckets to use in the brute force phase in order to achieve a better precision;

– *IFV Prec.*/ *IFV SMAPE* — precision and *SMAPE* metrics of the faiss *IFV*;

– *SSIFV Prec.*/ *SSIFV SMAPE* — precision and *SMAPE* metrics of the novel semi-supervised *IFV* proposed in the paper.

Table 1

***X*-size**	***Q*-size**	***K***	***N*probe**	***IFV Prec.***	***IFV SMAPE***	***SSIFV Prec.***	***SSIFV SMAPE***
10*K*	10*K*	200	1	0.055	8.7%	0.083	7.7%
10*K*	10*K*	200	5	0.200	4.37%	0.255	3.69%
10*K*	10*K*	200	20	0.480	1.81%	0.524	1.56%
1*M*	10*K*	2000	1	0.063	7.41%	0.071	7.1%
1*M*	10*K*	2000	5	0.200	3.8%	0.220	3.72%
1*M*	10*K*	2000	20	0.435	1.79%	0.491	1.65%

$X \sim N(0,1)$; $Q \sim N(0,1)$ results

Table 2

***X*-size**	***Q*-size**	***K***	***N*probe**	***IFV Prec.***	***IFV SMAPE***	***SSIFV Prec.***	***SSIFV SMAPE***
10*K*	10*K*	200	1	0.057	8.68%	0.066	8.53%
10*K*	10*K*	200	5	0.197	4.40%	0.207	4.33%
10*K*	10*K*	200	20	0.473	1.87%	0.460	1.95%
1*M*	10*K*	2000	1	0.061	8.16%	0.069	7.99%
1*M*	10*K*	2000	5	0.218	4.32%	0.217	4.34%
1*M*	10*K*	2000	20	0.490	1.77%	0.498	1.77%

$X \sim Exponential\ (1)$; $Q \sim Exponential(1)$ results

Table 3

***X*-size**	***Q*-size**	***K***	***N*probe**	***IFV Prec.***	***IFV SMAPE***	***SSIFV Prec.***	***SSIFV SMAPE***
10*K*	10*K*	200	1	0.025	14.76%	0.137	3.87%
10*K*	10*K*	200	5	0.107	6.14%	0.403	1.46%
10*K*	10*K*	200	20	0.305	2.49%	0.756	0.41%
1*M*	10*K*	2000	1	0.035	11.68%	0.141	3.65%
1*M*	10*K*	2000	5	0.130	4.97%	0.419	1.28%
1*M*	10*K*	2000	20	0.341	2.44%	0.766	0.40%

$X \sim N(0,1)$; $Q \sim Exponential(1)$ results

CONCLUSION

The experimental results of our novel semi-supervised modification to the Inverted File (IVF) index approach for approximate nearest neighbor search look very promising, because SS-IVF approach outperforms the industry standard implementation in a lot of different experiment configurations from the raw precision/smape metrics perspective, especially in scenarios where query distribution significantly differs from the indexed dataset. However, this SS-IVF algorithm is still quite far from a production solution, since we have not yet done an efficient C/C++ implementation, which would use parallelization and low-level optimizations.

REFERENCES

1. P. Indyk and R. Motwani, “Approximate nearest neighbors: towards removing the curse of dimensionality,” in *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, 1998.
2. H. Jégou, M. Douze, and C. Schmid, “Product Quantization for Nearest Neighbor Search,” *IEEE Xplore*. [Online]. Available: https://ieeexplore.ieee.org/ document/5432202
3. G. Voronoi, “Une méthode géométrique pour la détermination des régions de visibilité dans le voisinage d’un point de l’espace (A geometric method for determining regions of visibility in the vicinity of a point in space),” *Journal de Mathématiques Pures et Appliquées (Journal of Pure and Applied Mathematics)*, 1908.
4. J. Johnson, M. Douze, and H. Jégou, “Optimizing Product Quantization for Nearest Neighbor Search,” *IEEE Xplore*. [Online]. Available: https://ieeexplore.ieee.org/ document/6619223
5. D. E. Rumelhart and J. L. McClelland, “Learning Internal Representations by Error Propagation,” *IEEE Xplore*. [Online]. Available: https://ieeexplore.ieee.org/ document/6302929
6. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*. 2016. Available: https://www.deeplearningbook.org/
7. X. Glorot and Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*. 2010. [Online]. Available: https://proceedings.mlr.press/v9/ glorot10a/glorot10a.pdf
8. D.P. Kingma, *Adam: A Method for Stochastic Optimization*. 2014. [Online]. Available: https://arxiv.org/abs/1412.6980
9. *PyTorch*. [Online]. Available: https://pytorch.org/
10. *faiss::IndexIVF Class Reference*. [Online]. Available: https://faiss.ai/cpp_api/struct/ structfaiss_1_1IndexIVF.html

Received 05.09.2023

INFORMATION ON THE ARTICLE

Anton A. Bazdyrev, ORCID: 0000-0001-8191-897X, Educational and Research Institute for Applied System Analysis of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine, e-mail: bazdyrev.anton@gmail.com

ПІДХІД З НАПІВКЕРОВАНИМ НАВЧАННЯМ В ІНВЕРТОВАНОМУ ФАЙЛОВОМУ ІНДЕКСІ ДЛЯ ПОШУКУ НАБЛИЖЕНОГО НАЙБЛИЖЧОГО СУСІДА / А.А. Баздирев

Анотація. Запропоновано удосконалення підходу з використанням інвертованого файлового індексу для пошуку наближених найближчих сусідів з використанням напівкерованого навчання та навчання з учителем з метою підвищення ефективності проміжної кластеризації та досягнення більш збалансованих розмірів кластерів. Запропонований метод полягає у створенні кластерів за допомогою нейронної мережі з розв’язанням завдання класифікації векторів запитів у той самий кластер, що і їхні відповідні найближчі сусідні вектори у вихідному наборі даних. У поєднанні з мінімізацією стандартного відхилення розмірів кластерів процес індексування стає більш ефективним і точним під час наближеного пошуку найближчих сусідів. Через емпіричну оцінку на тестовому наборі даних продемонстровано, що запропонований підхід до індексу виявився більш точним порівняно з індустрійно-стандартною реалізацією із фіксованими параметрами, включаючи загальну кількість кластерів та кількість кластерів, що виділяються для запитів. Метод перспективний для підвищення ефективності пошуку найближчих сусідів у великорозмірних наборах даних у різних застосуваннях, таких як інформаційний пошук, пошук за природною мовою, рекомендаційні системи тощо.

Ключові слова: пошук наближених найближчих сусідів, інвертований файловий індекс, дані високої розмірності, машинне навчання.

ПРОБЛЕМНО І ФУНКЦІОНАЛЬНО
ОРІЄНТОВАНІ КОМП'ЮТЕРНІ СИСТЕМИ
ТА МЕРЕЖІ

UDC 62-50
DOI: 10.20535/SRIT.2308-8893.2023.4.06

RAISING THE INFORMATION SECURITY AWARENESS AMONG SOCIAL MEDIA USERS IN THE MIDDLE EAST

HEND KHALID ALKAHTANI

Abstract. Social media presents both opportunities and risks for any firm. The Internet has recently made everything possible. Due to its low cost and rapid speed, it is in high demand. Due to the virtual technique of interacting through various social media apps like Instagram, WhatsApp, Twitter, Facebook, etc., people are drawn to social networking. Despite the fact that it offers advantages on both sides, new threats are constantly emerging. Social media usage is widespread, but awareness is low, which makes significant cyberattacks more likely. Numerous threat categories put consumers at risk for cyber security. This research reviewed literature on educating Middle Eastern social media users about information security. Additionally, this research examines various threats made via social media, offers countermeasures, and considers various detection methods.

Keywords: security, security awareness social media, WhatsApp, Twitter, Facebook.

INTRODUCTION

Nowadays, the Internet has become a social environment that includes community, value and norm [1]. Sites like YouTube, Facebook, Twitter, and Instagram have seen a huge increase in the number of users. With the advancement of technology, social network has become pervasive and used like never before. In fact, social media is a collection of websites and applications designed and its goal is to allow people to share the content they want in a fast, efficient, and real-time manner [2]. It is also an online digital communication tool with which you can share links, SMS messages, photos and videos, it can be accessed anytime and anywhere. In fact, there are many social media sites where a large number of people spend a long time using them. Moreover, the number of OSNs is increasing year by year [3].

Facebook was the first social media network in the list of the most popular social networks with a large number of accounts and nearly 2.6 billion monthly active users. It also carries a huge amount of information due to this large number of users [4]. Nevertheless, this wide spread may cause great harm to users' private information because it may facilitate access to and violation of this information, because users cannot choose and specify their own privacy preferences in applications [5]. This poses significant privacy risks by making users' private data available to applications when they are often not fully aware of the risk of disclosing such information [2]. Moreover, Internet technology inherently leads to security problems, cybercrime, hackers, and intruders. In fact, the characteristics of the Internet reinforce the network structures that may lead to the occurrence of major Internet theft and fraud which is referred to as cybercrime.

Social media users need awareness and knowledge regarding the importance of personal information security, known as Information Security Awareness (ISA). ISA focuses on how an individual is aware of information security policies, rules, and guidelines [6]. Furthermore, ISA can shape individual characteristics to be more interested in revealing self-information in the context of social media. Thus, in this review, we will talk about the level of information security awareness among social media users in the Middle East [1]. Privacy violation is one of the main problems faced by social media users. It presents an ongoing risk to these users.

LITERATURE REVIEW

Due to its extensive use in the most prominent industries including education, healthcare, and entertainment, social media in the Middle East has grown to play a significant role in our lives. The popularity of various online social media platforms like Twitter, YouTube, Facebook and other social networking applications has increased because of this growth in the social networking field [7]. Therefore, the study's literature review will cover the knowledge gap of the significant risk to the personal information post on social media platforms.

The Perspective of Social networking privacy

People in general enjoy exchanging private information with each other, at the same time they have an obvious lack of awareness and knowledge about what might happen if they do so voluntarily or how to stop illegal disclosure of their personal information [8]. Social media's structure encourages its users to contribute willingly by exposing personal information. Users may reveal their personal information if they believe the benefits outweigh the drawbacks.

Social media provides a platform for studying business trends, consumer opinions, trend-setting, and political movements. Online activities that lack enough knowledge and social security and privacy can lead to extreme catastrophes, such as electronic hacking in which personal and private information is required for harmful purposes. Popular social media platforms like Facebook and Twitter, for example, also have their own unique techniques for determining the social characteristics of its users without having to ask them directly [8].

Despite the security measures taken by social media producers and programmers to protect the user information, it is still possible for the personal data to fall into the wrong hands and be exploited. This issue has already occurred in the early months of 2018 known as The Facebook-Cambridge Analytica data scandal, which involved Cambridge Analytica consulting company using millions of Facebook users' personal information without their knowledge or agreement for political advertising [9].

To sum up, threats are growing daily despite the existence of numerous preventative strategies, especially that many social media platforms have the option of making the user's profile available to the public. Moreover, without the user's awareness, attackers and online hackers can have access to the user's private information, as well as analyze and use them at remarkable speeds intending to cause harm.

Information security awareness

Information security awareness is a process that modifies and changes users' attitudes toward safe information standards as well as their values, behaviors, practices, work habits, and organizational culture. Changing these standards and practices helps every user to recognize the information security policies, guidelines, and procedures that should be followed in order to avoid any electronic harassment. Thus, users are required to understand both general and personal information security concepts [10].

Self-disclosure is defined as any knowledge about oneself that is voluntarily and consciously shared with others. Social media users need to be familiar with the significance of protecting personal information. Furthermore, security awareness can alter a person's personality so that they are more concerned about exposing personal information in the social media platforms [9].

Furthermore, the weakest link in any business is its regular users, who receive very little security awareness training as organizations grow their usage of cutting-edge security technologies and continually train their security personnel. As a result, organized hackers are working very hard right now to develop cutting-edge hacking techniques that can be used to steal both personal data and money from the general population [14]. Additionally, the Middle East is a desirable target for cybercriminals due to the region's rapid internet use rate and low consumer security awareness of the threats that may arise [10].

Accordingly, to escalate information security awareness of users, threats and harmful activities by online hackers should be pointed out and acknowledged.

Threats in Social Media Platforms

There are various threats done on social media platforms and applications, which a large number of users are not aware about, including the following.

Multimedia content threats. Multimedia content include threats associated with static links, Video and audio conferencing, and steganography.

Static links are used by 48.6% of social networking service users to exchange information from interactive media. This act causes the exposure of personal information and data loss for the users.

Moreover, most social media users post their own video and audio content to social media platforms in order to share their skills and ideas, some people abuse these audios and videos by modifying them to make them uncomfortable or life threatening [12].

Another threat of multimedia is steganography, which concentrates on encoding secret communications in a form that only the sender can comprehend in order to conceal sensitive information in visual form without the recipient awareness. For instance, a car image can contain some sort of computer viruses that deletes or steals the users' system files [13].

Traditional threats. Traditional threats include digital stalking, spamming, phishing data, and click jacking. Digital stalking is the practice of following and stalking someone online via email, or through other electronic communication channels. Typically, stalking requires a person engaging in persistent annoying or threatening actions [12].

While spamming refers to unsolicited texts or emails, which are distributed with multiple copies over the internet. Usually, spam messages are about commercial advertising. Spamming consumes a significant amount of network capacity in addition to wasting people's time.

Phishing is a fraud strategy used to obtain private data by misrepresenting a reliable organization, such as a password. Attackers frequently utilize phishing emails to spread risky links [8]. Figure 1 shows an illustration of the phishing technique used by hackers.

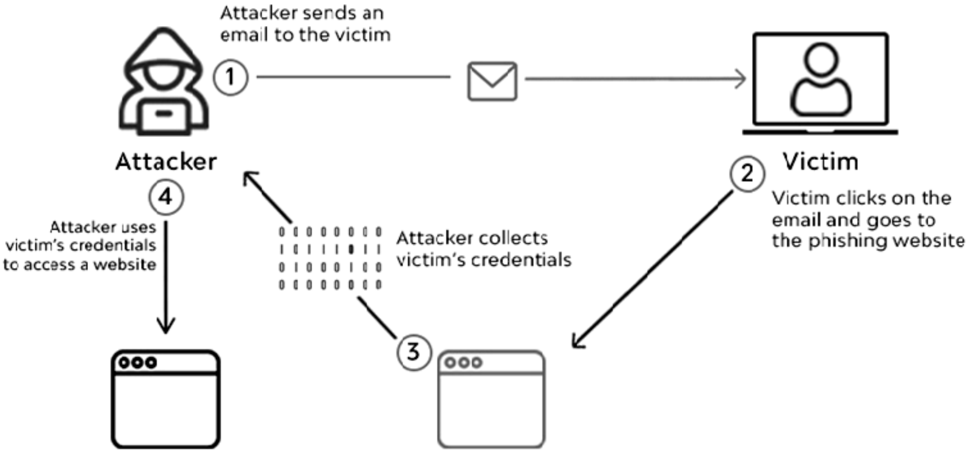

Attacker sends an email to the victim
1
Attacker
Victim
2
Victim clicks on the email and goes to the phishing website
4
Attacker uses victim's credentials to access a website
Attacker collects victim's credentials
3

Fig. 1. Phishing technique [11]

Another method of traditional threats is click jacking, which is based on deceiving a user into clicking on a wrong link. Users may unintentionally download spywares and viruses as a result, browse dangerous websites, or make unwanted online purchases.

Social Media Issues and Security Awareness among the users

Social media and its user have a relationship that is affected by cybersecurity and its environment. This relationship intensifies as social media use rises, as cybercriminals broaden their scope of interest and begin to focus on social media accounts. Users of social media are increasingly becoming one of their targets because the majority of them are unaware of the security and privacy measures that can be applied to individual accounts. Nowadays, social media plays a significant role in how individuals live their daily lives. This amply demonstrated the rise in popularity of social media in the modern day as well as the fact that its users have attained a critical mass necessary for influence, which increases their susceptibility and makes them even more vulnerable as hacking victims [15].

Teenagers in particular are using the Internet more and more frequently as a result of its increased popularity globally [16]. The majority of these adolescent social media users are students whose academic and social lives have been greatly impacted by changes in the global environment. These young adults and teenagers have no idea how to use their privacy settings. Teenagers and young adults want to express their identity and take the chance of being discovered and coming into touch with hackers because they are more interested in seizing the opportunity to connect with others and forge genuine relationships [17].

The majority of social media users do not really understand the significance of their privacy settings. Young adults and teenagers are more likely to be careless with their social media privacy settings. According to Livingstone's research findings from 2008, teens mostly use social media to create dangerous and intimate content by expressing themselves. Identity theft is one type of cybercrime that might occur as a result of this lack of privacy settings. The percentage of social media users who actively utilize various social network sites raises along with the overall growth of social media users. This brings us to the second action, where people are making it easier for hackers to find them. Users frequently link their social network account authorizations together or use the same password for several accounts because they maintain multiple social networks for personal usage [18].

For the hackers, this is a target straight out of heaven. Simply by acquiring access to one of the person's many accounts, they can quickly gain access to multiple accounts. A social network aggregator is what this is. Although it makes it easier for users to keep an eye on their social media profiles, it poses certain security risks [19]. Given that once one of their accounts has been hijacked, hackers will be able to find all of their other accounts, thereby increasing the risk to other accounts. For these teenagers and young adults, a cyber-security knowledge gap might be a problem. Due to their ignorance of the significance of security implementation, they are blind to nearby cases of accounts being compromised. Although social engineering assaults may not appear to be as sophisticated as other hacking techniques, they have produced some of the most effective attacks on targets [20].

The Need for Effective Information Security Awareness

Over the past few years, the Middle East has seen a steady rise in the number of internet users. While the Middle East only accounts for 3.2% of all internet users globally, it has seen an increase in internet usage of 1825% over the previous 10 years, compared to a rise of 445% for the rest of the globe, according to the World Internet Usage Statistics News [21]. Additionally, it stated that as of June 30, 2010, Bahrain, the United Arab Emirates, and Qatar had the greatest rates of internet penetration in the Middle East, representing 88%, 75.9%, and 51.8% of their respective populations, respectively. Numerous online businesses have been drawn to the Middle East by this expansion, enabling many already-established industries including education, health, aviation, and government to expand [22].

A thorough investigation of the difficulties and dangers that social networking sites and social networks face is the goal of Yassein M. et al. in [15]. In this study, electronic crimes were analyzed in relation to user-posted content on social networking sites and the use of that information to locate the original victims. Users are unaware of the risks associated with sharing this information when it is being published. After that, they list the flaws and give a brief summary of the protective strategies now in use, highlighting the weaknesses.

In [23] Almarabeh et al. discuss the distribution of the different sorts of attacks that social media sites are subject to. They present two different sorts of attacks in this context: classic attacks and modern attacks. A clear picture of the attacks has emerged thanks to the information on the different sorts of strikes. What kinds are there? And what techniques do they employ? There is information concerning social media users' flaws, such as the fact that their personal accounts have a low level of secrecy, which makes it simple to hack them. The primary goal of this research is to increase online social networking users' awareness of how to protect themselves and their data against risks and assaults while using social media platforms.

Security and privacy issues with social networks and social engineering were examined by Ali et al. [24]. Additionally, information about OSNs was covered, including its explanation, methods, resources, and the growth and fall of different OSNs. The report emphasized crucial privacy safeguards as well as user risks and weaknesses. By classifying risks into several categories that were investigated as part of a knowledge-sharing strategy, taxonomy has been constructed. In this study, the aftermath of a catastrophe is discussed, as well as how terrorism undermines the support for privacy. They incorporated privacy rules to take care of user privacy for the goal of reducing and monitoring personal information or data, encouraging users to reply appropriately and utilize social media solely for public topics.

The issues that could endanger the privacy of social media users were discussed by Ali et al. in [25]. There were two categories of issues: traditional threats and contemporary threats. To learn more about users' attitudes on privacy settings as well as their knowledge and interest in them, a questionnaire was created. Unfortunately, the findings were disappointing because a significant portion of users did not take use of privacy-preserving settings offered by service providers. Finally, recommendations and fixes for safeguarding user content and privacy were made. In [26] Aghasian et al. introduced an automated Fuzzy model for calculating the score of privacy of unstructured data of social media users on Facebook and Twitter. They also cautioned these users of the risks associated with using social media. The model consists of two phases that record privacy and calculate the privacy risk score. The machine learning model first identifies the features that have an impact on users' privacy. The final privacy score is then calculated using a fuzzy based approach. Information retrieval and pre-processing make up the first of the model's three phases. The second is by giving them some fundamental information so they can obtain the source of sentimental privacy. The outcome of the privacy score for users who have shared their information is determined at the final stage (Fig. 3).

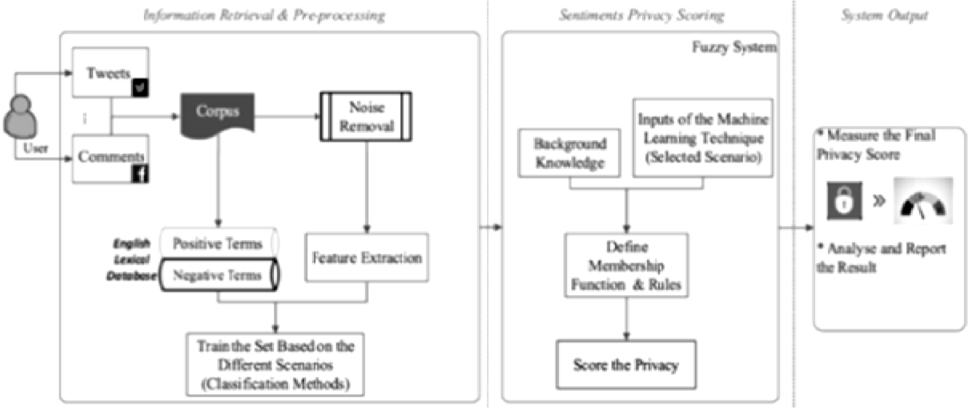

Information Retrieval & Pre-processing
Sentiments Privacy Scoring
System Output
Fuzzy System
User
Tweets
Comments
Corpus
Noise Removal
English Lexical Database
Positive Terms
Negative Terms
Feature Extraction
Train the Set Based on the Different Scenarios (Classification Methods)
Background Knowledge
Inputs of the Machine Learning Technique (Selected Scenario)
Define Membership Function & Rules
Score the Privacy
* Measure the Final Privacy Score
»
* Analyse and Report the Result

Fig. 3. Flow of privacy scoring process of proposed method [24]

Soomro et al. talked about a new list of crimes in social media like cyber intrusion, credit card fraud alongside disaster fraud and data breaches in [27], as well as many different types of crimes pertaining to social media like social engineering and phishing, burglary via social networking, identity theft, malware, cyber-casing, and cyber-stalking. Researchers in this area have employed a variety of strategies and deterrents to combat these crimes. Finally, this study emphasized a number of tips and methods for preventing cybercrime.

A deep learning-based efficient categorization technique for assaults that target Twitter users was proposed by Mostafa et al. [28]. The sole supporting method, the feature extraction issue, the lack of precision, and the slow speed were all issues this paper addressed. The tweets are first pre-processed using Sen2Vec rather than the element being extracted. The method used in this work is a sophisticated deep learning language processing technique that can convert a word or piece of text into a vector that represents it. Then, using a variety of machine learning methods, a machine learning model is created to distinguish between spam and non-spam. At the following level, parameter settings are assigned for spam filtering. An actual ground truth dataset is used to build up their testing.

Users were informed about the issues and potential harms caused by the dissemination of content on social media sites and the absence of privacy by, [26], and [28]. In this instance, the most significant security flaws affecting social media platforms and the most significant contemporary methods preventing the spread of hazardous content were investigated. Users and their exploitation or targeting were discussed. In order to lessen users' lack of privacy on social networking sites and to stop the spread of hazardous content in social networks, [24] offered a cutting-edge technology that has been researched and developed. "Fuzzy-based" is the name of the technology in use.

CONCLUSION

Social media is effective and beneficial in many areas, but new challenges and concerns regarding privacy and security continue to grow. This review discussed the most important papers that talk about the problems caused by the content of social networking sites, and the review also dealt with information security among social media users in the Middle East. This issue remains a fundamental and important issue. Accordingly, research and studies are ongoing. We suggest for future studies to focus more on information security awareness among social media users in the Middle East.

As Middle Eastern organizations expand their use of social media, advanced security technology, and use of the latest hardware and software, launching technical attacks has become more and more difficult. Similarly, organizations develop complete and well-written security policies and hire IT security experts who also help reduce the number of potential attacks. Unfortunately, little is used to secure the weakest link, that is, social media users. This drives attackers to gain unauthorized access to information by exploiting the user's trust and propensity to help. The paper discussed the level of information security awareness among social media users in the Middle East and reported the results of several IT security awareness studies. Discuss the importance of assessing security awareness by conducting monitoring audits. Several key factors have also been shown to help raise security awareness among social media users.

REFERENCES

1. L. Zhang, C. Amos, and I. Pentina, "Information Disclosure on a Chinese Social Media Platform," *J. Inf. Priv. Secur.*, vol. 11, no. 1, pp. 3–18, 2015. doi: 10.1080/15536548.2015.1010981
2. S. Rathore, P.K. Sharma, V. Loia, Y.S. Jeong, and J.H. Park, "Social network security: Issues, challenges, threats, and solutions," *Information Sciences*, 421, pp. 43–69, 2017. doi: 10.1016/j.ins.2017.08.063.
3. M. Al-Enazi and S. El Khediri, "Advanced Classification Techniques for Improving Networks Intrusion Detection System Efficiency," *Journal of Applied Security Research*, 17(1), 2021. doi: 10.1080/19361610.2021.1918500.
4. A. Ali, A. Kamran, M. Ahmed, B. Raza, and M. Ilyas, "Privacy concerns in online social networks: A users' perspective," *International Journal of Advanced Computer Science and Applications*, 10(7), 2019.
5. S. Ali, N. Islam, A. Rauf, I.U. Din, M. Guizani, and J.J. Rodrigues, "Privacy and security issues in online social networks," *Future Internet*, 10(12), pp. 1–12, 2018. doi: 10.3390/fi10120114.
6. M. Koohikamali, D.A. Peak, and V.R. Prybutok, "Beyond self-disclosure: Disclosure of information about others in social network sites," *Comput. Human Behav.*, vol. 69, pp. 29–42, 2017. doi: 10.1016/j.chb.2016.12.012.
7. F. Aloul, "The Need for Effective Information Security Awareness," *Journal of Advances in Information Technology*, 3(3), pp. 176–183, 2012. doi: 10.4304/jait.3.3.176-183.
8. S. Alotaibi, K. Alharbi, H. Alwabli, H. Aljoaey, B. Abaalkhail, S. El Khediri, "Threats, crimes and issues of privacy of users' information shared on online social networks," 2021 *International Symposium on Networks, Computers and Communications (ISNCC)*. doi: 10.1109/ISNCC52172.2021.9615815.
9. Dony Martinus Sihotang et al., "Factors Affecting the Intention of Social Media Users to Disclosure Personal Information," *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. doi: 10.1109/ICACSIS53237.2021.9631351.
10. A. Ali, M. Mahmud, N. Molok, and Sh. Talib, "Information security awareness through the use of social media," *The 5th International Conference "Information and Communication Technology for The Muslim World"*, 2014. doi: 10.1109/ICT4M.2014.7020668.
11. "Cyber security Brief Guide for Beginners: Phishing Attacks", *Cyber Coastal*, 2022. Available: https://cybercoastal.com/cybersecurity-brief-guide-for-beginners-phishing-attacks/
12. A. Gupta, Sh. Mehetre, and A. More, "Security in Social Media," *International Research Journal of Innovations in Engineering and Technology*, 5(12), pp. 40–44, 2021. doi: 10.47001/IRJIET/2021.512008.
13. Y. Hafsari, F. Permatasari, and N. Rahman, "A Review on Social Media Issues and Security Awareness among the users," *Journal of Applied Technology and Innovation*, 1(1), pp. 28–36, 2017.
14. P. Potgieter, "The Awareness Behavior of Students on Cyber Security Awareness by Using Social Media Platforms: A Case Study at Central University of Technology," *Proceedings of 4th International Conference on the Internet, Cyber Security and Information Systems 2019*, 12, pp. 272–280. Available: https://doi.org/10.29007/gprf
15. "Global Digital Statistics. GWI Social Summary. GWI Quarter Report," *Global Web Index*. Accessed on: February 1, 2017. [Online]. Available: http://insight. globalwebindex.net/hsfs/hub/304927/file-2377691590-pdf/Reports/GWI_Social_Summary_Q4_2014.pdf?submissionGuid=d75c46ce922c-4efc-8ac3-08dbe9ba4904
16. S. Bennett, A. Bishop, B. Dalgarno, J. Waycott, and G. Kennedy, *Implementing Web 2.0 technology in higher education: A collective case study.* 1st ed., 2012
17. A. Charlesworth, *An introduction to social media marketing*. 1st ed. London: Routledge, 2015, 209 p.

18. K. Lewis, J. Kaufman, and N. Christakis, "The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network," *Journal of Computer-Mediated Communication*, 14(1), pp.79–100, 2008. Available: https://doi.org/10.1111/j.1083-6101.2008.01432.x
19. F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user navigation and interactions in online social networks," *Information Sciences*, 195, pp. 1–24, 2012. doi: 10.1145/1644893.1644900.
20. N.A. Abd Rahman, "A Review on Social Media Issues and Security Awareness among the users," *Journal of Applied Technology and Innovation*, vol. 1, no. 1, pp. 28–36, 2018.
21. "Miniwatts Marketing Group," *2010 Internet World Stats*. Available: http://www.internetworldstats.com/stats.htm
22. F. Aloul, "The Need for Effective Information Security Awareness," *Journal of Advances in Information Technology*, 3, pp. 176–183, 2012. doi: 10.4304/jait.3.3.176–183.
23. H. Almarabeh and A. Sulieman, "The impact of cyber threats on social networking sites," *International Journal of Advanced Research in Computer Science*, 10(2), 2019. doi: 10.26483/ijarcs.v10i2.6384.
24. S. Ali, N. Islam, A. Rauf, I.U. Din, M. Guizani, and J.J. Rodrigues, "Privacy and security issues in online social networks," *Future Internet*, 10(12), 2018. doi: 10.3390/fi10120114.
25. E. Aghasian, S. Garg, and J. Montgomery, "An automated model to score the privacy of unstructured information—Social media case," *Computers & Security*, 92(3), 101778, 2020. doi: 10.1016/j.cose.2020.101778.
26. T.R. Soomro and M. Hussain, "Social Media-Related Cybercrimes and Techniques for Their Prevention," *Applied Computer Systems*, 24(1), pp. 9–17, 2019. doi: 10.2478/acss-2019-0002.
27. M. Mostafa, A. Abdelwahab, and H.M. Sayed, "Detecting spam campaign in twitter with semantic similarity," in *Journal of Physics: Conference Series*, vol. 1447, no. 1, p. 012044, 2020.
28. K. Stokes and N. Carlsson, "A peer-to-peer agent community for digital oblivion in online Social networks," in *2013 Eleventh Annual Conference on Privacy, Security and Trust, IEEE 2013*, pp. 103–110.

Received 15.12.2022

INFORMATION ON THE ARTICLE

Hend Khalid Alkahtani, College of Computer and Information Sciences of Princess Nourah Bint Abdulrahman University, Saudi Arabia, e-mail: hkalqahtani@pnu.edu.sa

ПІДВИЩЕННЯ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ СЕРЕД КОРИСТУВАЧІВ СОЦІАЛЬНИХ МЕДІА НА БЛИЗЬКОМУ СХОДІ / Хенд Халід Алкахтані

Анотація. Соціальні медіа створюють як можливості, так і ризики для будь-якої компанії. Інтернет нещодавно зробив все можливим. Завдяки низькій вартості і швидкості він користується великим попитом. Завдяки віртуальній техніці взаємодії через різні програми соціальних мереж, такі як Instagram, WhatsApp, Twitter, Facebook тощо, людей приваблює використання соціальних мереж. Незважаючи на те, що Інтернет пропонує переваги для обох сторін, постійно виникають нові загрози. Соціальні мережі використовуються широко, але поінформованість низька, що підвищує ймовірність значних кібератак. Існує багато категорій загроз, які загрожують споживачам. У праці розглянуто літературу про навчання користувачів соціальних мереж Близького Сходу щодо інформаційної безпеки. Крім цього, розглянуто різні загрози через соціальні мережі, запропоновано заходи протидії та розглянуто різні методи виявлення.

Ключові слова: безпека, соціальні медіа, WhatsApp, Twitter, Facebook.

UDC 62-50
DOI: 10.20535/SRIT.2308-8893.2023.4.07

NOVEL MODIFIED KERNEL FUZZY C-MEANS ALGORITHM USED FOR COTTON LEAF SPOT DETECTION

PRADIP M. PAITHANE, SARITA JIBHAU WAGH

Abstract. Image segmentation is a significant and difficult subject that is a prerequisite for both basic image analysis and sophisticated picture interpretation. In image analysis, picture segmentation is crucial. Several different applications, including those related to medicine, facial identification, Cotton disease diagnosis, and map object detection, benefit from image segmentation. In order to segment images, the clustering approach is used. The two types of clustering algorithms are Crisp and Fuzzy. Crisp clustering is superior to fuzzy clustering. Fuzzy clustering uses the well-known FCM approach to enhance the results of picture segmentation. KFCM technique for image segmentation can be utilized to overcome FCM's shortcomings in noisy and nonlinear separable images. In the KFCM approach, the Gaussian kernel function transforms high-dimensional, nonlinearly separable data into linearly separable data before applying FCM to the data. KFCM is enhancing noisy picture segmentation results. KFCM increases the accuracy rate but ignores neighboring pixels. The Modified Kernel Fuzzy C-Means approach is employed to get over this problem. The NMKFCM approach enhances picture segmentation results by including neighboring pixel information into the objective function. This suggested technique is used to find "blackarm" spots on cotton leaves. A fungal leaf disease called "blackarm" leaf spot results in brown leaves with purple borders. The bacterium can harm cotton plants, causing angular leaf blotches that range in color from red to brown.

Keywords: Cluster Accuracy Rate (CAR), Clustering, Cotton Leaf Disease, Fuzzy Clustering Method (FCM), Kernel Fuzzy C-means Algorithm (KFCM), Novel Modified Kernel Fuzzy C-Means Clustering Algorithm (NMKFCM).

INTRODUCTION

Cotton is the most significant cash crop farmed in Maharashtra, India. The primary issue reducing cotton output is disease on the plant. Because a minute difference in color pattern might be caused by a different disease that is present on a cotton leaf, we know that the human eye's perception is not powerful enough to enable it to recognize minute variations in the diseased region of an image. The cotton plant's leaf is the disease's primary source. The leaves of the cotton plant are where 80–90% of the illness is located. One crucial technique for separating a picture into its backdrop and its objects is image segmentation. Clustering is one of the crucial phases in picture segmentation. In the early portion of the season, affected crops may develop slowly or be stunted. Blackening of the roots is a symptom of the illness, which results in the destruction of the root cortex (outer layer). Thielaviopsis basicola does not kill seedlings on its own, however some roots may perish. Significant black root rot exposes the root to Pythium or Rhizoctonia infection. When growth begins in warmer temperatures, the dead cells of the root cortex are shed, and plants that were severely harmed earlier in the season may not continue to exhibit symptoms later in the season.

- Diseases on Leaves of Cotton.
- The diseases on the cotton leaves are classified as:
  - – Bacterial disease: e.g. Bacterial Blight, Crown Gall, Lint Degradation;
  - – Fungal diseases: e.g. Anthracnose, Leaf Spot;
  - – Viral disease: e.g. Leaf Curl, Leaf Crumple, Leaf Roll;
- Diseases like alternaria leaf spot, Bacterial blight, Bacterial stunt, Black root, Boll rot/tight lock.

The collection of observations is divided into smaller groups so that observations within each group are somewhat comparable to one another. Multivariate data analysis typically uses clustering as a routine practice. It is intended to investigate the data objects' innate natural structure, where items in the same cluster are as similar as possible to one another and objects in separate clusters are as distinct as possible from one another. The method used to arrange items or patterns so that samples from the same group resemble one another more than samples from other groups. There have been many different clustering techniques employed, including the hard clustering scheme and the fuzzy clustering scheme, each of which has unique particular traits. Each data point can only belong to one cluster when using the traditional hard clustering approach. As a result, when using this method, the segmentation results are frequently quite precise, meaning that every pixel in the image belongs to exactly one class. Yet, in many actual scenarios, problems with pictures like inadequate contrast, noise, overlapping intensities, and insufficient spatial resolution make this hard (crisp) segmentation a challenging process.

Types of Clustering:

- Hard: same object can only belong to single cluster.
- Soft: same object can belong to different clusters.

The current days, deep learning approach is used for cotton leaf segmentation. The CNN, VGG-16, VGG-19, ResNet-50 and some hybrid model has been used for this problem. The deep learning approach has been improved the accuracy of cotton leaf image segmentation as compared to state-of-art. The NMKFCM model is also gives stable result as compared to deep learning approaches. In deep learning approaches, training time period is major constraint for this problem. In the experimental analysis, the training time is near about 1 hour to 2 hour and in the proposed method the training process is not required.

MATERIAL AND METHODS

Fuzzy Clustering

In image segmentation, a soft segmentation technique has received extensive study and effective application. Since it has resilient qualities for ambiguity and can preserve significantly more information than hard segmentation methods, the Fuzzy C-Means (FCM) algorithm is the most widely used fuzzy clustering approach in picture segmentation. The typical FCM method has a severe flaw in that it lacks spatial context information, making it vulnerable to noise and imaging artefacts even while it performs well on the majority of noise-free pictures.

Fuzzy C-Means Algorithm. A well-liked and practical image division algorithm is FCM. The FCM algorithm was created by Dunn and enhanced by Bezdek [1]. This method is intended to scale back an objective goal [2]. Because each quality vector may only belong to one cluster and the quality vectors of the data set can be separated into solid clusters, this method outperforms the k-mean technique. Instead, the FCM loosens the restriction and enables the quality vector to assign a range of association scores to diverse clusters. Suppose a set of data with related clusters. A data value is equidistant from both clusters while also being near to them.

Activity in the clustering loop is FCM. By reducing the intragroup biased sum of the squared error task J_m function, it produces the best c partitions [3]:

$$J_m = \sum_{j=1}^{C} \sum_{i=2}^{N} U_{i,j}^m d_{i,j}^2 ,$$

where N — the number of patterns in X; C — the number of clusters; U_{ij} — the degree of membership; W_j — the center of cluster j; d_{ij} — distance between object X_i and cluster center W_j; m — the biased value.

The FCM algorithm focuses on minimizing J_m, subject to the following constraints on U:

$$U_{ij} \in [0,1], \ i = 1,2,3,\ldots,N, \ \text{and} \ j = 1,2,3,\ldots,C;$$

$$\sum_{j=1}^{C} U_{ij} = 1, \ \ i = 1,2,3,\ldots,N, \ \ 0 < \sum_{i=1}^{N} U_{ij} < 1, \ \ j = 1,2,3,\ldots,C.$$

Objective function J_m describe a constrained optimization problem, which can be converted to an unconstrained optimization problem by using Lagrange multiplier technique. By using this calculates membership function and update cluster center separately:

$$U_{ij} = \frac{1}{\sum_{i=1}^{c} \left(\frac{d_{ij}}{d_{il}} \right)^{\frac{2}{(m-1)}}}, \ \ i = 1,2,\ldots,N, \ \text{and} \ \ j = 1,2,\ldots,C;$$

$$\text{If } d_{ij} = 0 \ \text{ then } \ U_{ij} = 1 \ \text{ and } \ U_{ij} = 0 \ \text{ for } \ 1 \neq j .$$

And calculate cluster center using following step

$$w_j = \frac{\sum_{i=1}^{N} (U_{ij})^m x_i}{\sum_{i=1}^{N} (U_{ij})^m} \ \ , \ \ j = 1,2,\ldots,C .$$

The FCM algorithm focuses on minimizing objective function Jm. It fails in noisy image to detect accurate and sharp image segmentation process.

Kernel Fuzzy C-Means Algorithm. The FCM algorithm calculates the distance between the cluster center and the data item using Euclidian distance. FCM fails in noisy and nonlinear data sets because Euclidian distance does not perform as intended in noisy data. Kernel Fuzzy C-means technique is used to address this flaw. Kernel information is used with FCM in the KFCM approach [4]. KFCM works by mapping input data into a higher-dimensional feature space and utilizing the Kernel technique to transform nonlinearly separable data into linearly separable data. While using the kernel approach, the data set was complicated and nonlinear before becoming simple and separable when using the FCM method [5].

KFCM classifies noisy objects into clusters with greater clarity than FCM and with greater accuracy in noisy images. The value of the KFCM membership matrix U may range from 0 to 1

KFCM is iterative clustering methods that generate optimal c partition by using minimize objective function J_{kfcm}:

$$J_{km}(U,W) = 2\sum_{j=1}^{C}\sum_{i=1}^{N} U_{ij}^{m}(1 - K(X_i, W_j)) .$$

In this objective function Gaussian kernel function is used:

$$K(X,Y) = \exp\left(-\frac{x - y^2}{\sigma^2}\right).$$

In KFCM clustering algorithm choose initial cluster randomly and perform following step.

1. Provide Gaussian kernel function for input image.
2. Evaluate membership function between object and cluster center.
3. Evaluate new updated cluster center.
4. Repeat step iteratively until no new cluster found.

KFCM it work properly in noisy image but KFCM not focus on neighborhood term.

Modified Kernel Fuzzy C-Means (MKFCM). This method is intended to scale back an objective goal [6]. Because each quality vector may only belong to one cluster and the quality vectors of the data set can be separated into solid clusters, this method outperforms the k-mean technique. Instead, the Fuzzy C-Mean loosens the restriction and enables the quality vector to assign a range of association scores to diverse clusters. Suppose a set of data with related clusters. A data value is equidistant from both clusters while also being near to them.

FCM is looping clustering activity. It generates optimum c partitions by abating the intragroup biased sum of the squared error task J_m function [7].

Kernel Method. The kernel methodology is a method that, by replacing the internal product with an appropriate Mercer Kernel, generates an implicit non-line map of the feedback information to a high-level quality space [8]. The kernel may be used in any method that solely depends on the dot product between two vectors. Every time a kernel is applied, a dot product is replaced. When two data are planned into a high-level-dimensional space, the space metrical that calculates the space between them is simplified. It is easier to tell apart and more distinctly differentiated [9].

Feature Space Mapping. Consider a non-line map task $\Phi : I = \mathbb{R}^2 \to F = \mathbb{R}^3$ from the 2-dimensional input space *I* into the 3-dimensional feature space *F*:

$$\varnothing(\vec{x}) = (x_1^2, \sqrt{2x_1x_2}, x_2^2)^T \,. \tag{1}$$

Hyperplane is represented by Eq. for separable dataset:

$$\vec{w}^T \vec{x} + b = 0\,. \tag{2}$$

Consider the splitting hyperplane Eq. (1) into a linear task in $\mathbb{R}^3$:

$$\vec{w}^T \varnothing(\vec{x}) = w_1 x_1^2 + w_2\sqrt{2x_1x_2} + w_3 x_2^2 = 0\,.$$

Eq. (2) is an elliptic job when as usual value of a constant *c* and assessed in $\mathbb{R}^2$.

In Fig. 1 any nonlinear separable data is converted into linear separable, so every pixel is classified on the basis of a feature.

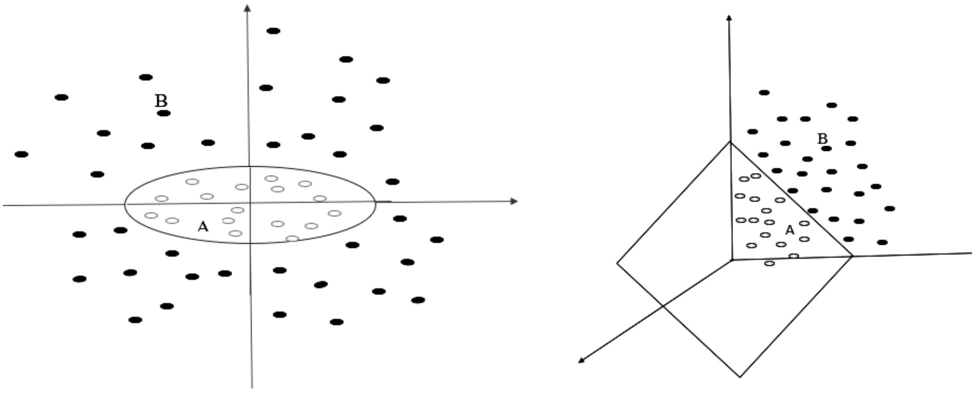


Fig. 1. Conversion of Non-line Distinguishable Data into Line Distinguishable Data

Use the appropriate mapping function to use *F*'s linear classifier with the converted form of the data to find a non-straight classifier without hassle. After mapping the non-line distinguishable data to a high-level space, *I*, locate a hyperplane that distinguishes linearly. For sensitive learning consider Fig. 1.

It depends only on the data mapped by the inner product of the feature space *F*. Defining a function $K(\vec{x}_i, \vec{x}) = \varnothing(\vec{x}_i)^T \varnothing(\vec{x})$, called kernel, that directly calculates the dot product of the mapping data places in the quality space eliminates the need for even the explicit coordinates of *F* or the mapping task [10]. The subsequent standard sample of a kernel "*K*" shows the computation of the dot product in the quality space applying $K(\vec{X}, \vec{Z}) = (\vec{X}^T, \vec{Z})^2$. It is encouraging the map task $\Phi(\vec{x}) = (x_1^2, \sqrt{2}x_1, x_2, x_2^2)^T$:

$$\vec{x} = (x_1, x_2)\,, \quad \vec{z} = (z_1, z_2)\,;$$

$$K(\vec{X}, \vec{Z})(\vec{X}^T, \vec{Z})^2 = (x_1 z_1 + x_2 z_2)^2 =$$

$$= (x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2) = (x_1^2, \sqrt{2x_1x_2}, x_2^2)^T (z_1^2, \sqrt{2z_1z_2}, z_2^2)\,.$$

The advantage of such a kernel operation is that the complexness of the improvement of drawback continues solely reliant on the spatial property of the "*input space*" and not of the "*quality space*".

Different types of Kernels are mentioned below [11]:

Linear Kernel function: $K(x,z) = x^T z$;

Polynomial Kernel function: $K(x,z) = (x^T z + \theta)^d$;

Gaussian Kernel function: $K(x,z) = \exp\left(-\frac{x - z^2}{\sigma^2}\right)$;

Sigmoid Kernel function: $K(x,z) = \tanh(\alpha((x^T z) + \theta))$.

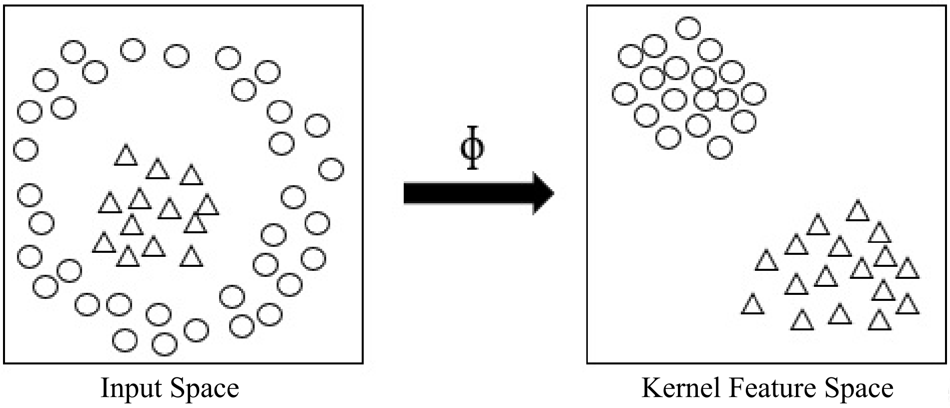


Fig. 2. Kernel Feature Space

Figure 2 is depicted the kernel working process for liner separable process to detect correct segmented regions. NMKFCM method is integrating closer pixel quantity in objective function [12]. NMKFCM method is a revised form of KFCM. KFCM is unsuitable for images damaged by instinct disturbance. KFCM has operated accurately in indistinct and nonlinear separable data, but it doesn't consist of information of closer pixel, to overcome this drawback, introduced NMKFCM is integrating closer pixel cost by applying "3×3" or "5×5" window window. A closer pixel quantity is included in objective task [13; 14]. Thec "α" constraint is applied to manage the impact of closer's term. It is having upper cost with growth of image disturbance. Scale of α cost rests within "0 to 1", if ratio of disturbance is minimal then take cost of α between "0 and 0.5". Ratio of disturbance is above average then take cost of α is "0.5 and 1.0". It is a beneficial and useful algorithm as compared to other algorithms. It has achieved sharp outcomes in disturbance images.

It is a looping procedure. It reduces the cost of objective tasks through closer pixel. In this objective task, present window across pixel and "α"parameter [15]:

$$J_{NMKFCM_{obj}}(\mathrm{U},\mathrm{W}) = \sum_{y=1}^{Q}\sum_{x=1}^{P} U_{xy}^{m}(1 - \mathrm{K}_T(\mathrm{Z}_x, \mathrm{W}_y))\left(\frac{N_R - \alpha \sum\limits_{k\in N_i} U_{yk}}{N_R}\right), \quad (3)$$

where N_R — the cardinality; N_i — set of closer pixel value include into a window across pixel Z_i. Objective task j_{nmkn} illustrate a constrained optimization dilemma. Eq. 3 is applied for conversion into an unconstrained optimization dilemma. In Eq. 3, Lagrange multiplier technique is used.

By applying this computes membership function and update cluster center separately:

$$U_{xy}=\left(\frac{\left((1-\mathrm{K}_T(\mathrm{Z}_x,\mathrm{W}_y))\left(\frac{N_R-\alpha\sum_{l\varepsilon N_i}U_{yl}}{N_R}\right)\right)^{-\frac{1}{m-1}}}{\sum\left((1-\mathrm{K}_T(\mathrm{Z}_x,\mathrm{W}_y))\left(\frac{N_R-\alpha\sum_{l\varepsilon N_l}U_{kl}}{N_R}\right)\right)^{-\frac{1}{m-1}}}\right).$$

And calculate cluster center using following step:

$$W_y=\frac{\sum_{y=1}^{P}U_{xy}^{m}K_T(\mathrm{Z}_x,\mathrm{W}_y)\,\mathrm{Z}_x}{\sum_{x=1}^{Q}U_{xy}^{m}(\mathrm{Z}_x,\mathrm{W}_y)}.$$

Algorithm: Objective Function of NMKFCM

```
INPUT
    1. $Z=\{Z_1,Z_2,\ldots,Z_N\}$, Data set
    2. $P$, $2\le P\le y$, $y$ is number of cluster
    3.Define cost of $\mathcal{E}$, used to terminate loop
    4.Set membership function $U_{xy}^{0}$ using input data and cluster.
    5.Determine cluster center $W_0=(w_{01},w_{02},\ldots,w_{0p})$
OUTPUT
    $W_j=\{W_0,W_2,\ldots,W_p\}$, targeted center of clusters.
begin
    for
      $t$=0
    if $\{U^t-U^{t+1}\}<\varepsilon$
        Update center $W_p^t$ $with$ $U^t$ by using Eq.
        Update membership function $U^{t+1}$ by using Eq.
        $t$+1
    else
        segmented output
    end
```

This method is advantageous to integrate closer pixel information. Standard FCM and IFCM methods are responsive to disturbance and preliminary cluster centers. It is ignoring the 3-D correlation of pixels, leading to inaccurate clustering outcomes [16]. NMKFCM work very fit in neighborhood pixel material.

Goal and Objectives:

- Choosing value of alpha to improve accuracy of image segmentation.

- Add Gaussian kernel method and RBF function to give more accuracy and also work in noisy and noiseless.
- Determine required number of clusters for image segmentation.
- Improving CAR value in all image formats.

Hills Climbing Algorithm

Image segmentation is a crucial step in the processing of images. Applications like image segmentation, adaptive compression, and region-based image retrieval benefit from the detection of conspicuous picture areas. Saliency is measured by comparing an image region's local contrast to its surrounding area at different scales [17]. It is using a contrast determination filter that runs at various scales to produce saliency maps with saliency values per pixel for the purpose of identifying salient locations. These separate maps come together to form the final saliency map [18]. We employ a rather straightforward segmentation approach to show how the final saliency map may be used to segment whole objects [19].

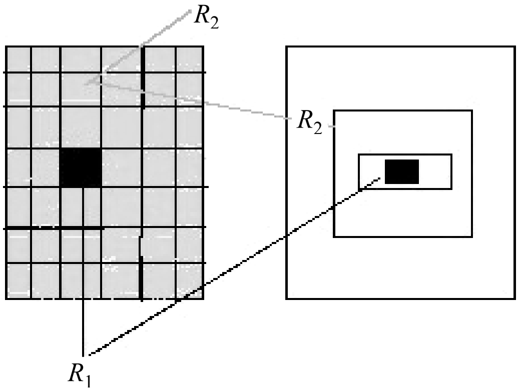


Fig. 3. Saliency map with R_1 inner and R_2 outer region

In this, compare the distance between the average feature vectors of the pixels in a subregion of the picture with the pixels in the area around it. Instead of merging separate saliency maps for scalar values of each feature, this enables the creation of a combined feature map at a particular scale utilizing feature vectors for each pixel [20]. The distance D between the average vectors of pixel characteristics of the inner area R_1 and that of the outer region R_2 is what determines the contrast-based saliency value $c(i,j)$ for a pixel at location $I(j)$ in the picture:

$$c_{i,j} = D\left[\left(\frac{1}{N_1}\sum_{p=1}^{N_1} v_p\right),\left(\frac{1}{N_2}\sum_{q=1}^{N_2} v_q\right)\right],$$

where v is the vector of feature elements corresponding to a pixel and N_1 and N_2 are the number of pixels in R_1 and R_2, respectively. If v is a vector containing uncorrelated feature items, then the distance D is a Euclidean distance; if the vector's elements are correlated, then the distance D is a Mahalanobis distance. In this study, feature vectors for color and brightness are generated using the CIELab color space and RGB photographs. Although the CIELab colour space's perceptual differences are roughly Euclidian, D in equation [13]:

$$c_{i,j} = v_1 - v_2 ,$$

where $v_1 = [L_1; a_1; b_1]T$ and $v_2 = [L_2; a_2; b_2]T$ are the average vectors for regions R_1 and R_2, respectively.

The final saliency map is determined as a sum of saliency values across the scales S:

$$m_{i,j} = \sum_{s} c_{i,j}\,.$$

The hill-climbing technique may be thought of as a search window that is ran through the d-dimensional histogram's space to locate the biggest bin inside of it. Each bin in the colour histogram has $3d - 1 = 26$ neighbors because the CIELab feature space is three-dimensional, where d is the number of dimensions in the feature space. The values of these bins serve as the starting seeds, and the number of peaks obtained reveals the value of K [21]. By adding up values in the final saliency map M that correspond to pixels in the segmented picture, the average saliency value V per segmented region is determined:

$$V_k = \frac{1}{|r_k|} \sum_{i,j \in r_k} m_{i,j}\,,$$

$|r_k|$ is the segmented region's size in pixels. The segments with an average saliency value greater than a predetermined threshold T are maintained, while the other segments are removed, according to a straightforward threshold-based procedure. As a consequence, the output only includes the segments that make up the salient item.

The $L^*a^*b^*$ color space enables us to quantify these differences. The $L^*a^*b^*$ color space is derived from the CIE XYZ tristimulus values. The $L^*a^*b^*$ space comprises of a luminosity layer 'L^*', chromaticity-layer 'a^*' indicating where color falls along the red-green axis, and chromaticity-layer 'b^*' indicating where the color falls along the blue-yellow axis [22].

Algorithm: Hill-climbing Based Segmentation.

Input: An Image.

Output: a group of aesthetically connected segments.

1. Create the image's color histogram.

2. Ascend the color histogram's slope from a non-zero bin to the apex as shown below:

2.1. The amount of pixels in the current histogram bin should be compared to the numbers in the adjacent (left and right) bins.

2.2. The algorithm moves upwards towards the neighboring bin with the greater number of pixels if the surrounding bins have differing amounts of pixels.

2.3. The algorithm checks the next nearby bins if the immediate neighbors have the same amount of pixels, and so on, until two neighboring bins with different numbers of pixels are discovered. Next, a shift upward is performed to the bin with the most pixels.

2.4. Repeat steps 2.1–2.3 to continue going upwards until you reach a point from which you can travel no further uphill. When the adjacent bins contain less pixels than the current bin, that is the situation. As a result, the present bin is considered a high.

Choose a different unclimbed bin to use as your starting bin, then follow step 2 to locate another summit. This process is repeated until the color histogram's non-zero bins are all climbed (associated with a peak). The discovered peaks are preserved since they indicate the input image's original number of clusters.

2.5. The halting bin is designated as the peak of a hill if no upward progress is made, and all bins going to this peak are connected to it.

3. Choose a different unclimbed bin to use as your starting bin, then follow step 2 to locate another summit. This process is repeated until the histogram's non-zero bins are all climbed (associated with a peak).

4. The recognized peaks are preserved because they show how many clusters there were in the input picture at the beginning.

5. The same peak's neighboring pixels are clustered together.

Lastly, pixels that are close to one another and lead to the same peak are grouped together, assigning each pixel to a different peak. Hence, create the input image's clusters.

EXPERIMENTAL RESULT

Evaluation Parameter

1. Cluster Accuracy Rate

$$CAR = \left| \frac{A \cap S}{A \cup S} \right| .$$

2. Dice

$$dice(A,S) = 2 * \frac{|A \cap S|}{\| A| + |S\|} .$$

3. IOU

$$IOU = \frac{|A \cap S|}{|A \cup S|} .$$

4. Bfscore

$$bfscore = \frac{2 * precision * recall}{(recall + precision)} ,$$

where $A =$ output image; $S =$ input image.

Detail comparison of proposed method with traditional method (see Table 1–3)

Table 1. Detail Comparison of Proposed Method with Traditional Method

Image Name	Approach	Evaluation Parameter		
		IOU	bfscore	dice
Image 1	FCM	55.78	33.51	71.62
	KFCM	73.81	26.36	84.93
	NMKFCM	81.55	41.25	89.83
Image 2	FCM	68.61	21.28	81.38
	KFCM	84.47	36.81	91.58
	NMKFCM	89.88	42.29	98.81

Continued Table 1

Image Name	Approach	Evaluation Parameter		
		IOU	bfscore	dice
Image 3	FCM	90.22	37.14	94.86
	KFCM	80.66	34.71	89.29
	NMKFCM	90.75	36.98	95.15
Image 4	FCM	72.94	18.11	84.35
	KFCM	90.06	40.33	94.77
	NMKFCM	94.51	54.46	97.18
Image 5	FCM	74.95	26.51	85.68
	KFCM	73.17	25.37	84.51
	NMKFCM	80.36	27.42	87.99

Table 2. Detail Comparison of Proposed Method with Traditional Method

Image Name	Approach	Cluster Accuracy Rate(CAR)
Image 1	FCM	63.71
	KFCM	71.78
	NMKFCM	74.98
Image 2	FCM	57.9021
	KFCM	64.8723
	NMKFCM	69.9572
Image 3	FCM	67.2396
	KFCM	64.3482
	NMKFCM	70.2246
Image 4	FCM	58.6222
	KFCM	66.3623
	NMKFCM	69.184
Image 5	FCM	86.8277
	KFCM	86.0188
	NMKFCM	95.337

Table 3. Detail comparison of proposed method with traditional method

Image Name	Approach	Time Period
Image 1	FCM	12.24
	KFCM	10.44
	NMKFCM	8.47
Image 2	FCM	14.24
	KFCM	08.37
	NMKFCM	06.54
Image 3	FCM	11.61
	KFCM	11.29
	NMKFCM	09.24
Image 4	FCM	13.19
	KFCM	12.66
	NMKFCM	07.29
Image 5	FCM	12.24
	KFCM	08.59
	NMKFCM	07.76

The Table 4 depicts the detail comparison of NMKFCM approach with deep learning approaches. The NMKFCM is providing strong and stable result as compare to CNN model. The CAR value of NMKFCM approach is 98.80 which higher than other approaches. The IOU and Precision value of NMKFCM achieved higher result value as compared to deep learning models. The NMKFCM is having less value The Dice value of NMKFCM improvised the result as compare to state-of-art. in bfscore as compare to other approaches.

Table 4. Detail comparison of proposed method with traditional method

Approach	CAR	IOU	Precision	Dice	bfscore	Time	Training Time
CNN [23]	95.37	92.00	0.8750	46.0	87.50	8~9 second	94 Minute
VGG16 [23]	98.10	92.18	0.9583	46.0	95.16	5~6 second	54 Minute
ResNet-50 [23]	98.32	91.49	0.9482	50.0	95.65	2~3 second	53 Minute
Menon Model[23]	98.53	94.23	0.9579	50.0	96.42	5~6 second	77 Minute
NMKFCM	98.80	94.51	1.0000	98.81	54.46	4~6 second	Not Required

The execution time of NMKFCM is less than CNN, VGG-16 and Menon Model, but higher than ResNet-50. The training time is not required for NMKFCM (Fig. 4).

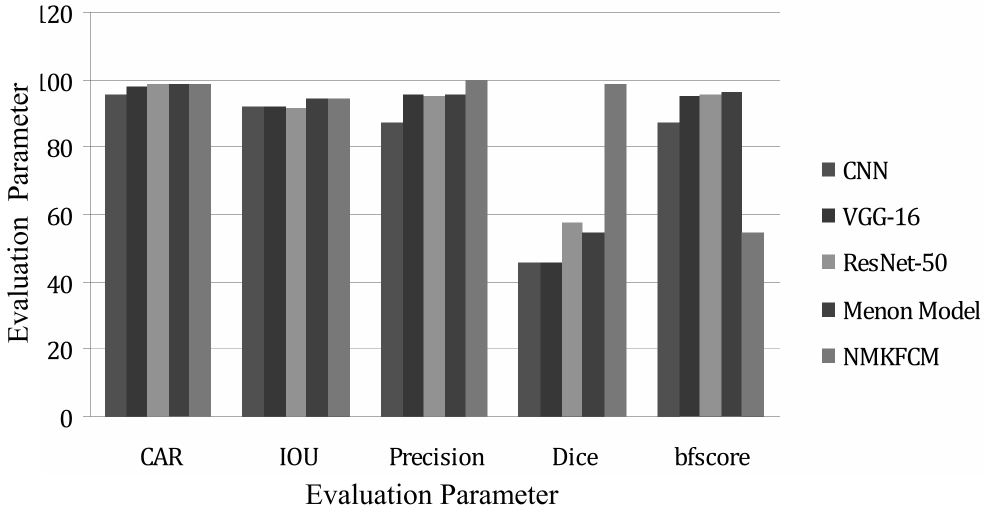


Fig. 4. Comparison of NMKFCM with Deep Learning Approaches

In above image, sub image (A), (E), (I), (M) and (Q) are original image of cotton leaf. Sub image (B), (F), (J), (N) and (R) are segmented by FCM approach, Sub image (C), (G), (K), (O) and (S) are segmented by KFCM approach, Sub image (D), (H), (L), (P) and (T) are segmented by NMKFCM approach. The sub image (A) is affected by Bacterial Blight disease, The sub image (E) is affected by Leaf Curl, The sub image (I) is affected by alternaria leaf spot , The sub image (M) is affected by fungal disease.

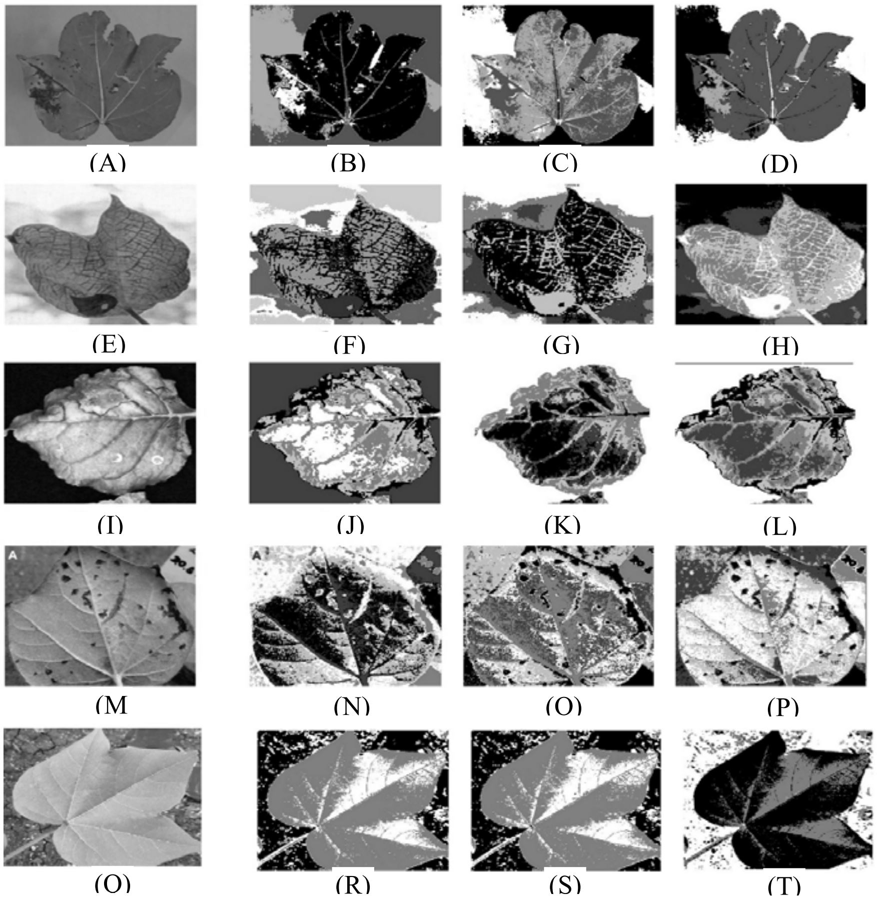


Fig. 5. Cotton Leaf Image Segmentation using FCM, KFCM and NMKFCM

CONCLUSION

The crucial and indispensable element of an image analysis system, image segmentation is a key area of study for many image processing researchers. Four methods — Clustering, Thresholding, Region Extraction, and Edge Detection — are used to segment images. Clustering is the downgrouping of related data elements. Here, we've used techniques for clustering like crisp and fuzzy. In this system, Fuzzy C-Means, Kernel Fuzzy C-Means, and Modified Kernel Fuzzy C-Mean Clustering are all used as clustering techniques. In comparison to FCM and Crips clustering methods, MKFCM is a suggested system that provides accurate picture segmentation while also enhancing segmentation performance by adding the influence of neighbor pixel information. The MKFCM method can automatically identify the necessary cluster number for picture segmentation with the use of the Hill climbing algorithm. The suggested technique can automatically estimate the cluster number for a noisy picture, but this number is not helpful for image segmentation since the proposed algorithm has formed a cluster for noisy pixels, making image segmentation less effective than for noiseless pixels. In the future, we will be able to select an alpha value to increase the precision of picture segmentation and CAR (Cluster Accuracy Rate) values across all image formats.

The proposed method is not required training time but in deep learning approaches training is mandatory. The proposed method is improvising the IOU, precision, Dice and CAR value as compared to deep learning approaches.

No conflict of interest.

REFERENCES

1. P.M. Paithane and S.N. Kakarwal, “Automatic determination number of cluster for multi kernel NMKFCM algorithm on image segmentation,” in *International Conference on Intelligent Systems Design and Applications*, pp. 870–879. Springer, Cham, 2018.
2. Chun-Yan Yu, Ying Li, Ai-lian Liu, and Jing-hong Liu, “A novel modified kernel fuzzy c- means Clustering algorithms on Image segmentation,” *2011 14th IEEE International Conference*. doi: 10.1109/CSE.2011.109.
3. Saiful Islam and Dr. Majidul Ahmed, “Implementation of Image Segmentation for Natural Images using Clustering Methods,” *IJETAE*, vol. 3, issue 3, March 2013.
4. L.A. Zadeh, “Fuzzy Sets”, *Information and Control*, 8, pp. 338–353, 1965.
5. Songcan Chan and Daoqiang Zhang, “Robust Image Segmentation Using FCM With Spatial Constraints Based on New Kernel Induced Distance Measure,” *IEEE transactions on Systems, MAN and Cybernetics-Part B: Cybernetics*, vol. 34, no. 4, August 2004.
6. P.M. Paithane, S.N. Kakarwal and D.V. Kurmude, “Automatic Seeded Region Growing with Level Set Technique Used for Segmentation of Pancreas,” in *International Conference on Soft Computing and Pattern Recognition*, pp. 374–382. Springer, Cham, 2020.
7. P.M. Paithane and S.N. Kakarwal, “Automatic Pancreas Segmentation using A Novel Modified Semantic Deep Learning Bottom-Up Approach,” *International Journal of Intelligent Systems and Applications in Engineering*, 10(1), pp. 98–104, 2022.
8. Pradip Mukundrao Paithane, “Yoga Posture Detection Using Machine Learning,” *Artificial Intelligence in Information and Communication Technologies, Healthcare and Education: A Roadmap Ahead*, 2022.
9. S. Kakarwal and Pradip Paithane, “Automatic pancreas segmentation using ResNet-18 deep learning approach,” *System Research and Information Technologies*, no. 2, pp.104–116, 2022.
10. Elnomery Zanaty and Sultan Aljahdali,“Improving Fuzzy Algorithms for Automatic Magnetic Resonance Image Segmentation,” *The International Arab Journal of Information Technology*, vol. 7, no. 3, pp. 271–279, July 2009.
11. Kaur Prabhjot, Gupta Pallavi, and Sharma Poonam, “Review and Comparison of kernel Based Fuzzy Image Segmentation Techniques,” *I.J. Intelligent Systems and Applications*, 7, pp. 50–60, 2012.
12. Robert L. Cannon, Jintendra V. Dave, and James C. Bezdek, “Efficient Implementation of the Fuzzy c-Means Clustering Algorithms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, issue 2, March 1986.
13. Martin Hofman, *Support vector Machines-Kernel and the Kernel Trick,* pp. 1–16, 2006.
14. Daoqiang Zang and Songcang Chen, “Fuzzy Clustering Using Kernel Method,” *Proceedings of the 2002 International Conference on Control and Automation,* Xiemen, China, June 2002. doi: 10.1109/ICCA.2002.1229535.
15. Daoqiang Zang and Songcang Chen, “A novel kernalized fuzzy C-means algorithm with Application in medical image segmentation,” *Artificial Intelligence in Medicine*, 32, pp. 37–50, 2004.
16. E.A. Zanaty, Sultan Aljahdli, and Narayan Debnath, “A Kernalized Fuzzy C-Means Algorithm for Automatic Magnetic Resonance Image Segmentation,” *Journal of Computational Methods in Sciences and Engineering Archive*, vol. 9(1,2S2), pp. 123–136, April 2009. doi: 10.3233/JCM-2009-0241.
17. Shailash Kochra and Sanjay Joshi, “Study on Hill-Climbing Algorithm for Image Segmentation Technology,” *International Journal of Engineering Research and Applications (IJERA*), vol. 2, issue 3, pp. 2171–2174, May-Jun 2012.
18. Garima Goyal, “TEM Color Image Segmentation using Hill Climbing Algorithm,” *International Journal of Computer Science and Information Technologies*, vol. 5, pp. 3457–3459, 2014.
19. A. Abirami Shri, E. Aruna, and Ajanthaa Lakkshmanan,“Image segmentation and recognition,” *International Journal of Computer Applications; 3rd National Conference on Future Computing, February 2014.*

20. Pradip M. Paithane, S.N. Kakarwal, and D.V. Kurmude, "Top-down method used for pancreas segmentation," *Int. J. Innov. Exploring Eng. (IJITEE)*, vol. 9, issue 3, pp. 1790–1793, 2020.
21. Sarita Jibhau Wagh, Pradip M. Paithane, and S.N. Patil, "Applications of Fuzzy Logic in Assessment of Groundwater Quality Index from Jafrabad Taluka of Marathawada Region of Maharashtra State: A GIS Based Approach," *Hybrid Intelligent Systems: 21st International Conference on Hybrid Intelligent Systems (HIS 2021), December 14–16, 2021*. Cham: Springer International Publishing, 2022.
22. Pradip M. Paithane and Sarita Jibhau Wagh, "Automatic Quality Control Scrutiny of Sugar Crystal using K-Means Clustering Algorithm Image Processing," *American Scientific Research Journal for Engineering, Technology*, and Sciences, 9(12):2395-0056, 2022.
23. M.S. Memon, P. Kumar, and R. Iqbal, "Meta Deep Learn Leaf Disease Identification Model for Cotton Crop," *Computers*, 11(7), 102, 2022. Available: https://doi.org/10.3390/computers11070102
24. Pradip Paithane, Sarita Jibhau Wagh, and Sangeeta Kakarwal, "Optimization of route distance using k-NN algorithm for on-demand food delivery," *System Research and Information Technologies*, no. 1, pp. 85–101, 2023. doi: 10.20535/SRIT.2308-8893.2023.1.07.

Received 01.04.2023

INFORMATION ON THE ARTICLE

Pradip M. Paithane, ORCID: 0000-0002-4473-7544, Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, India, e-mail: paithanepradip@gmail.com

Sarita Jibhau Wagh, ORCID: 0000-0003-4798-2147, T.C. College Baramati, India

НОВИЙ МОДИФІКОВАНИЙ АЛГОРИТМ ЯДРА FUZZY C-MEANS, ЩО ВИКОРИСТОВУЄТЬСЯ ДЛЯ ВИЯВЛЕННЯ ПЛЯМ НА ЛИСТУ БАВОВНИКА / Прадіп М. Пейтане, Саріта Джібхау Ваг

Анотація. Сегментація зображення є важливою та складною темою, яка є необхідною умовою як для базового аналізу зображення, так і для складної інтерпретації зображення. В аналізі зображень сегментація зображення має вирішальне значення. Кілька різних програм, зокрема ті, що стосуються медицини, ідентифікації обличчя, діагностики хвороби Коттона та виявлення об'єктів на карті, отримують переваги від сегментації зображення. Для сегментації зображень використовується підхід кластеризації. Існує два типи алгоритмів кластеризації: чіткий і нечіткий. Техніка чіткості перевершує нечітку кластеризацію. Нечітка кластеризація використовує добре відомий підхід FCM для поліпшення результатів сегментації зображення. Техніка KFCM для сегментації зображення може бути використана для усунення недоліків FCM у зашумлених і нелінійних роздільних зображеннях. У підході KFCM ядрова функція Гауса використовується для перетворення високовимірних нелінійно розділених даних у лінійно розділені дані перед застосуванням FCM до даних. KFCM поліпшує результати сегментації зображення із шумом, підвищує рівень точності, але ігнорує сусідні пікселі. Щоб подолати цю проблему, використовується модифікований підхід нечіткого С-середнього ядра. Підхід NMKFCM поліпшує результати сегментації зображення шляхом включення інформації про сусідні пікселі до цільової функції. Цей запропонований метод використовується для виявлення плям «чорної шкірки» на листу бавовника. Грибкове захворювання листя під назвою «чорна плямистість» призводить до коричневого листя з фіолетовими краями. Бактерія може завдати шкоди рослинам бавовника, спричиняючи кутасті плями на листу, які мають колір від червоного до коричневого.

Ключові слова: коефіцієнт точності кластера (CAR), кластеризація, хвороба листя бавовника, метод нечіткої кластеризації (FCM), алгоритм нечіткого С-середнього ядра (KFCM), новий модифікований алгоритм кластеризації нечіткого С-середнього ядра (NMKFCM).

UDC 004.942:614.4(460)
DOI: 10.20535/SRIT.2308-8893.2023.4.08

INFORMATION SYSTEM FOR ASSESSING THE INFORMATIVENESS OF AN EPIDEMIC PROCESS FEATURES

K. BAZILEVYCH, O. KYRYLENKO, Y. PARFENIUK, S. YAKOVLEV, S. KRIVTSOV, I. MENIAILOV, V. KUZNIETCOVA, D. CHUMACHENKO

Abstract. The primary objective of this study is to assess the informativeness of various parameters influencing epidemic processes utilizing the Shannon and Kullback–Leibler methods. These methods were selected based on their foundation in the principles of information theory and their extensive application in machine learning, statistics, and other relevant domains. A comparative analysis was performed between the results acquired from both methods, and an information system was designed to facilitate the uploading of data samples and the calculation of factor informativeness impacting the epidemic processes. The findings revealed that certain features, such as "Chronic lung disease," "Chronic kidney disease," and "Weakened immunity," did not carry significant information for further analysis and hindered the forecasting process, as per the data set examined. The developed information system efficiently supports the assessment of feature informativeness, thereby aiding in the comprehensive analysis of epidemic processes and enabling the visualization of the results. This study contributes to the current body of knowledge by providing specific examples of applying the described algorithmic models, comparing various methods and their outcomes, and developing a supportive tool for analyzing epidemic processes.

Keywords: information system, epidemic process, informativeness of features, Shannon method, Kullback–Leibler method.

INTRODUCTION

Predicting morbidity is an essential task in health care and public health. The use of machine learning in the analysis of epidemic processes is relevant in contemporary conditions, as it allows for rapid and efficient processing of large volumes of data and making accurate forecasts [1]. This helps reduce the consequences of epidemics and ensures a more effective fight against diseases. Using machine learning models helps predict morbidity with high accuracy [2].

In the modern world, especially considering the current situation related to the COVID-19 pandemic, the theme of analyzing data on epidemic processes remains extremely relevant and critically important. Data analysis is an essential tool that plays a key role and helps understand the spread of disease [3], identify trends [4], identify risk groups of the population [5], evaluate the effectiveness of control measures [6], imagine the scale of the problem [7], and predict the future development of epidemics [8]. It helps scientists, doctors, and relevant authorities make informed decisions and develop strategies for effective epidemic control [9].

It is also difficult to overestimate the importance of timely medical diagnostics in managing epidemic processes. Rapid and accurate disease diagnosis is a key factor for successful control and management of epidemics [10]. Ensuring timely diagnostics allows diagnosing and isolating sick people, starting treatment, taking necessary preventive measures and vaccination, and taking strategic steps to reduce the spread of the disease.

Laboratory tests are one of the main tools for medical diagnostics of epidemic diseases [11]. They allow for detecting the presence of a pathogenic agent, determining its characteristic properties, and establishing a diagnosis. For example, in the case of the COVID-19 pandemic, testing for the SARS-CoV-2 virus is crucial for detecting infected individuals, even when they do not show symptoms. This helps to take appropriate control measures and preventive strategies.

Many modern healthcare facilities have information systems for storing various medical data about patients' health, used by doctors for diagnosing pathological processes [12]. However, when analyzing medical data, identifying patterns, and extracting it, one faces the problem of dimensionality. The dimensionality of stored data, determined by the number of different features describing the patient's health status, is vast and sometimes reaches several tens and hundreds of indicators [13].

Evaluating informativeness is essential for analyzing epidemic process data, as it allows for determining the significance of various factors and relationships associated with diseases [14]. This helps to identify key factors affecting the spread of epidemics and make effective decisions regarding their prevention and treatment. Informativeness evaluation also helps detect complex relationships between different factors and determine which has the most significant impact on epidemic processes [15]. This allows for making more accurate predictions and effective decisions regarding epidemic response.

Therefore, the problem of reducing the dimensionality of the feature space and identifying the most informative features is a very relevant task of epidemic process data analysis.

The aim of the paper is to develop the information system for evaluation of the factors' informativeness for healthcare data.

Research is part of a complex intelligent information system for epidemiological diagnostics, the concept of which is discussed in [16, 17].

2. MATERIALS AND METHODS

2.1. Informativeness of features

The informativeness of a feature is an indicator of its significance or usefulness for solving a specific task or problem. This is an essential concept in many areas, including machine learning, statistics, signal processing, and many others [18]. The informativeness of features is assessed depending on their ability to classify or predict the target variable. More informative features have a greater impact on the model and provide more significant information for the separation or prediction of classes.

Diagnostic features are specific symptoms, indicators, or characteristics used to diagnose a disease, condition, or problem [19]. In medicine, diagnostic features help doctors determine a disease or condition based on examination, patient surveys, laboratory tests, examinations, images, and other studies. Diagnostic features may include such indicators:

- Physical symptoms: for example, pain, pulsation, swelling, bleeding, skin color change, etc.

- Behavioral symptoms: for example, nervousness, depression, irritation, inability to concentrate, sleep change, appetite change, etc.
- Laboratory results: such as cell count, hormone level, substance concentration in the blood or urine, or results of other analyses.
- Imaging: results of X-rays, CT scans, MRI, or other techniques that may show changes in the structure or function of organs.
- Anamnesis: information obtained from the patient about their medical history, symptoms, duration, and nature of the disease.
- Genetic research: determining the presence or absence of certain genetic mutations or variants.

2.2. Problem formulation of feature space reduction

The application of modern information technologies in medicine contributes to accumulating large volumes of medical data, which are stored and processed using medical information systems (MIS). These data contain medical knowledge that can be extracted and used for decision-making, such as diagnosing pathological processes [20]. The dimensionality of the stored data, defined by the number of different features describing the patient's health status, is vast and sometimes reaches several tens and hundreds of indicators. Therefore, the problem of reducing the dimensionality of the feature space and highlighting the most informative features is very relevant for MIS development.

Let Ω be a set of objects, and $X = \{x_1, x_{2,} \ldots, x_n\}$ be the finite set of quantitative features of these objects. For any object $\varpi \in \Omega$, its feature description $\{x_1(\varpi), x_2(\varpi), \ldots, x_n(\varpi)\}$ is known as a n-dimensional vector, where this vector's $\left(i-a\right)$-th coordinate equals the ($i-a$)-th feature's value. The set of feature descriptions of objects for a given sample of objects $A \subseteq \Omega$ is given as a matrix of size $|A| \times n$, a table “object – feature”. Let $I(Z)$ be the measure of informativeness of the subset of features $Z \subseteq X$, defined on A. It is necessary to select some subset $Z^* \subseteq X$ from all different subsets of the set *X*, such that

$$I(Z^*) = \max_{Z \in X} I(Z).$$

The task of features selection is computationally complex; as for $|X| = n$, a permutation of all different subsets $Z \subseteq X$ requires $O(2^n)$ time.

2.3. Kullback–Leibler Method

The Kullback–Leibler method is a statistical approach for measuring the divergence between two probability distributions. This method is popular in many fields, including statistics, machine learning, and information theory [21]. Using the Kullback–Leibler method, a measure is calculated that gauges the divergence between two distributions to assess the informativeness of a feature.

Typically, two distributions are input into the Kullback–Leibler method to evaluate the informativeness of features [22]: the distribution of data with the feature value considered and the distribution of data without considering the feature value. The method estimates the informativeness of the studied feature as a value ranging from 0 to 2. In this case, it is considered that the closer the informativeness measure $I(x)$ is to 2, the higher the informativeness of *x*, and conversely, the closer $I(x)$ is to 0, the lower the informativeness of x. The output of the

Kullback–Leibler method is a numerical estimate indicating the informativeness of the feature.

Algorithmic Model of the Kullback–Leibler Method

Step 1. Define the target input set (in this case, it is “Morbidity”).

Step 2. Calculate the probability of the event for each value in the target set: $Q(X) = n(X)/N$, where *n* is the number of cases X, and N is the total number of cases.

Step 3. Calculate the probability of the event for each value in the feature: $P(y) = n(y)/N$, where *n* is the number of cases y, and N is the total number of cases.

Step 4. Calculate the Kullback–Leibler divergence between the two sets *P* and *Q*. The Kullback–Leibler divergence, sometimes called relative entropy, is a measure of the difference between two probability distributions:

$$D(P,Q) = \sum_i P(i)\log_2(P(i)/Q(i)),$$

where *P(i)* is the joint probability of the event *X*-target set and *y*-feature, and *Q(i)* is the probability of the event of the target set.

Repeat *steps 3-4* for all values in the feature and calculate the overall Kullback–Leibler divergence.

Step 5. Calculate the overall informativeness of the feature.

Step 6. Evaluate the obtained results based on the magnitude of the informativeness of the feature. The higher the evaluation value, the more informative the feature.

Step 7. Select the features with the highest values as the most informative.

The algorithm of the model is shown in Fig. 1.

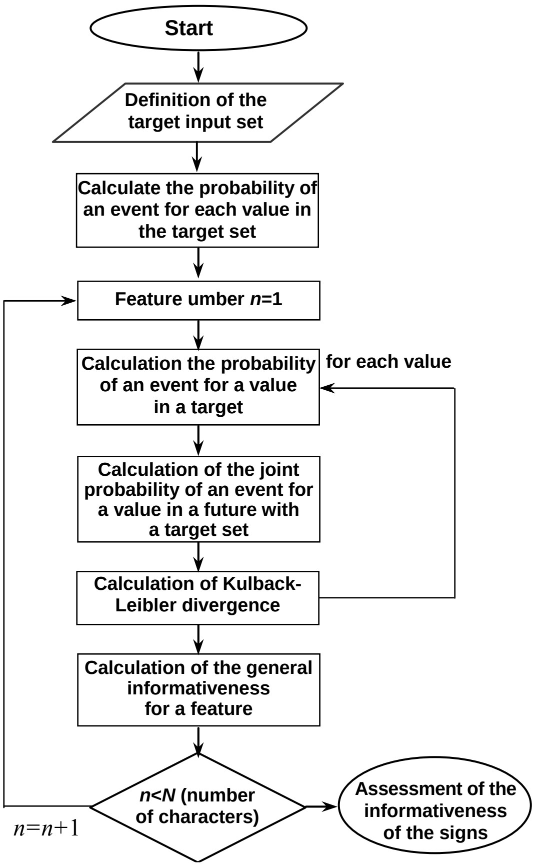

Start
Definition of the target input set
Calculate the probability of an event for each value in the target set
Feature umber *n*=1
Calculation the probability of an event for a value in a target
for each value
Calculation of the joint probability of an event for a value in a future with a target set
Calculation of Kulback-Leibler divergence
Calculation of the general informativeness for a feature
n<*N* (number of characters)
Assessment of the informativeness of the signs
n=*n*+1

Fig. 1. The algorithm of the Kullback–Leibler method

2.4. Shannon Method

The Shannon method for calculating feature informativeness in a table is based on the concept of entropy in information theory [23]. Entropy is a measure of uncertainty or randomness in a data set. Entropy reflects the average level of 'information,' 'surprise,' or 'uncertainty' inherent in the possible outcomes of a random variable [24].

The Shannon method provides an estimate of the informativeness of the studied feature in the form of a normalized variable, which takes values from 0 to *1* [25]. In this case, the informativeness of feature *x* is said to be higher as $I(x)$ approaches *1* and lower as $I(x)$ approaches 0.

Algorithmic model of the Shannon method

Step 1. Define the target input set (in our case, it is “Morbidity”).

Step 2. Calculate the total entropy for the target set using the Shannon formula

$$H(S) = -\sum_{i=0}^{N} p_i \log_2 p_i \ ,$$

where p_i is the probability of the occurrence of the *i*-th class in the data set, H is the entropy, and S is the set of instances.

Step 3. Divide the data by each unique feature value and calculate the frequency of each value in the target set.

Step 4. Calculate the entropy for each feature value.

Step 5. Calculate the weighted entropy for each feature value, multiplying the entropy value by its frequency. Weighted entropy by the Shannon method [26] is used to measure the informational weight of a random event:

$$H_{weighed} = P(S)H(S) \, ,$$

where $P(S) = m / N$: m is the frequency of the occurrence of the value in the feature; N is the total number; $P(S)$ is the probability of the occurrence of the *S*-th class relative to the target variable.

Step 6. Calculate the informativeness of features. The informativeness of a feature is calculated as the difference between the entropy of the output set and the sum of the entropy of the subsets formed by the given feature, with weights equal to the fraction of the subset in the output set:

$$I(S) = H(S) - \sum\nolimits_{i=0}^{N} H_{weighed} \ ,$$

where $I(S)$ is the informativeness of the feature of the subset S .

Repeat *steps 2-6* for all features and calculate the informativeness for each feature.

Step 7. Evaluate the obtained results based on the informativeness of the feature. The higher the evaluation value, the more informative the feature.

Step 8. Select features with the highest values as the most informative.

Figure 2 shows the flowchart of the algorithmic model.

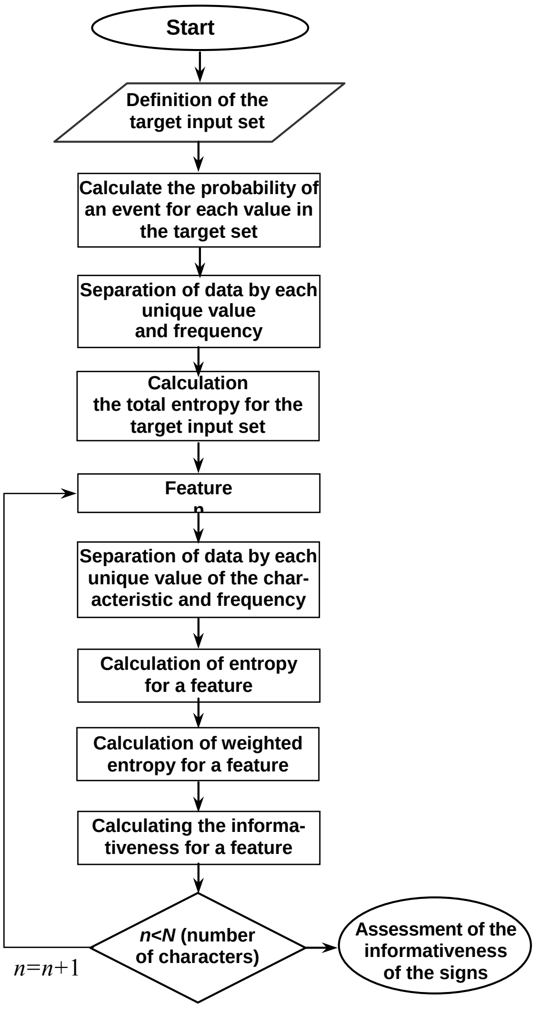

Start
Definition of the target input set
Calculate the probability of an event for each value in the target set
Separation of data by each unique value and frequency
Calculation the total entropy for the target input set
Feature n
Separation of data by each unique value of the characteristic and frequency
Calculation of entropy for a feature
Calculation of weighted entropy for a feature
Calculating the informativeness for a feature
n<N (number of characters)
n=n+1
Assessment of the informativeness of the signs

Fig. 2. The algorithm of the Shannon method

3. RESULTS

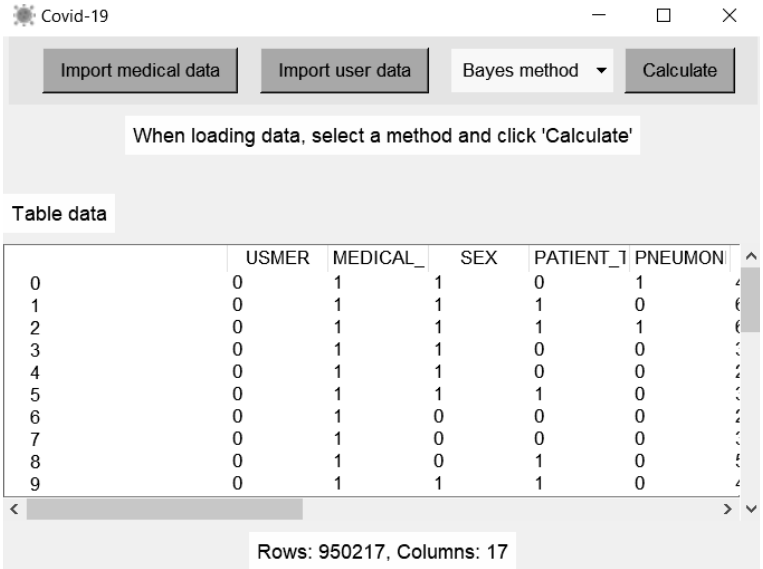
3.1. Program realization

Various algorithms and methods were employed to develop the information system, and *Python* is an ideal choice for such tasks. Its library, *sklearn*, includes many machine learning algorithms, including naive Bayes, logistic regression, and gradient boosting [27].

For *data visualization, tkinter, matplotlib.pyplot,* and *seaborn* were used, which are powerful visualization tools in Python. These libraries provide many possibilities for creating plots, diagrams, interactive visualizations, and more.

Based on data from *healthcare facilities*, the developed software product predicts the probability of a patient getting sick. The product is a decision-support system for general practitioners, which is especially important during pandemics and other disasters that limit the number of doctors.

Figure 3 shows the interface of the software product.



When loading data, select a method and click 'Calculate'

Table data

	USMER	MEDICAL_	SEX	PATIENT_T	PNEUMONI	
0	0	1	1	0	1	4
1	0	1	1	1	0	6
2	0	1	1	1	1	6
3	0	1	1	0	0	3
4	0	1	1	0	0	2
5	0	1	1	1	0	3
6	0	1	0	0	0	2
7	0	1	0	0	0	3
8	0	1	0	1	0	5
9	0	1	1	1	0	4

Rows: 950217, Columns: 17

Fig. 3. Decision support system interface

Further, by pressing the *"Calculate"* button, the calculation of informativeness estimation methods is carried out, precisely the Shannon method and the Kullback–Leibler method.

3.2. Data analysis

The experimental study used data on patients suffering from COVID-19 [28]. Figure 4 depicts the histogram of the input data.

Next, we checked the dataset for empty data that would worsen the prediction. Figure 5 shows all data output in terms of data type, presence of zero, and the number of records of 950217 patients.

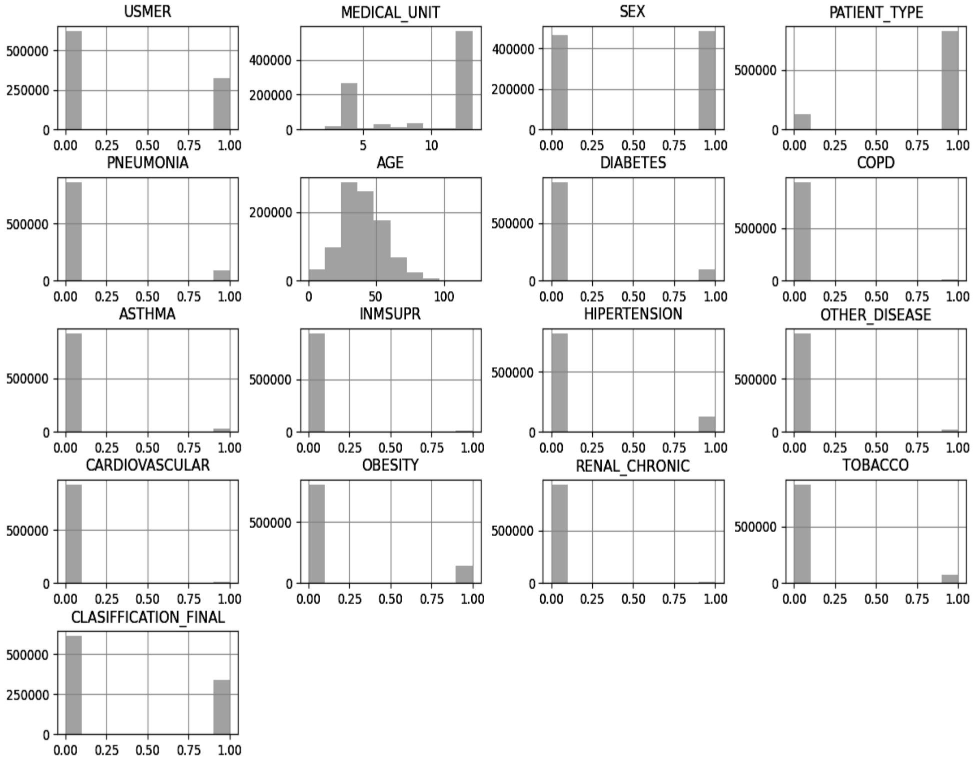
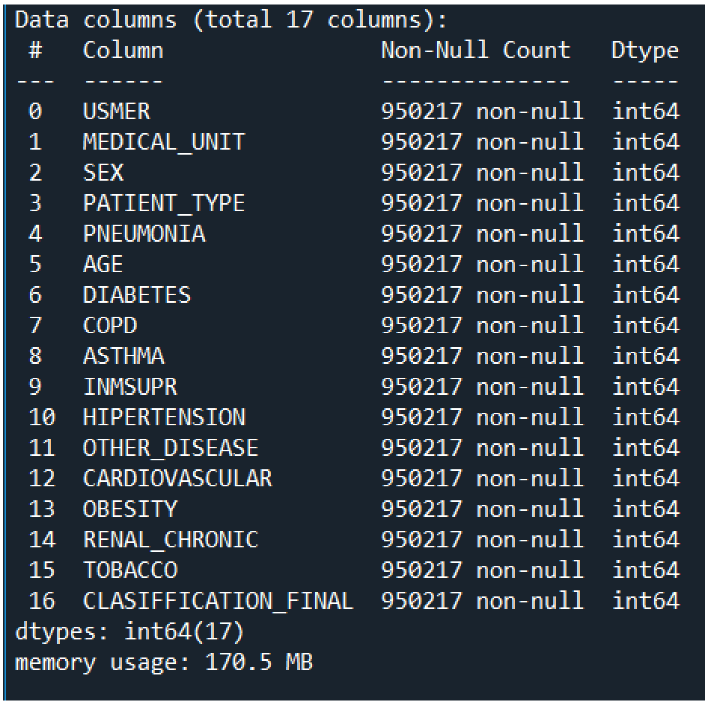


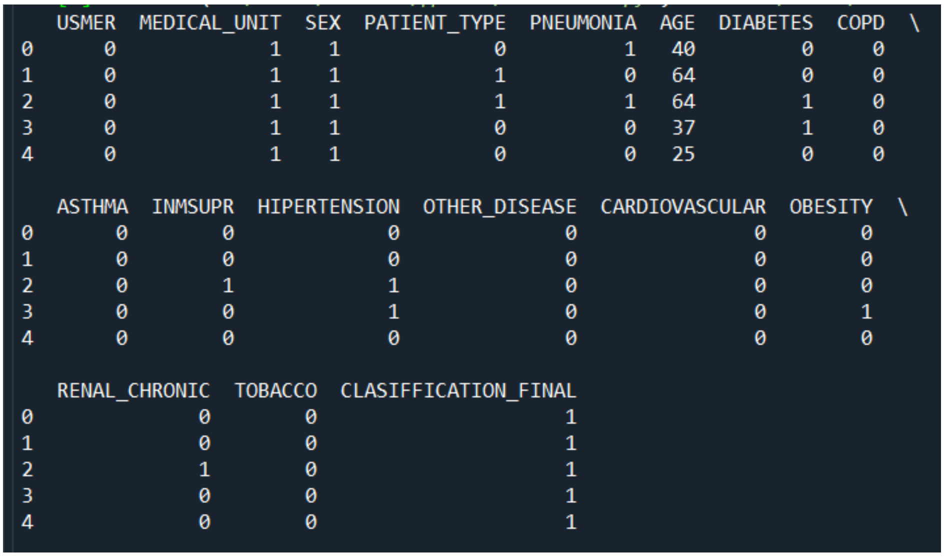
Fig. 4. Patient Data Histogram



```
Data columns (total 17 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   USMER                 950217 non-null  int64
 1   MEDICAL_UNIT          950217 non-null  int64
 2   SEX                   950217 non-null  int64
 3   PATIENT_TYPE          950217 non-null  int64
 4   PNEUMONIA             950217 non-null  int64
 5   AGE                   950217 non-null  int64
 6   DIABETES              950217 non-null  int64
 7   COPD                  950217 non-null  int64
 8   ASTHMA                950217 non-null  int64
 9   INMSUPR               950217 non-null  int64
 10  HIPERTENSION          950217 non-null  int64
 11  OTHER_DISEASE         950217 non-null  int64
 12  CARDIOVASCULAR        950217 non-null  int64
 13  OBESITY               950217 non-null  int64
 14  RENAL_CHRONIC         950217 non-null  int64
 15  TOBACCO               950217 non-null  int64
 16  CLASIFFICATION_FINAL  950217 non-null  int64
dtypes: int64(17)
memory usage: 170.5 MB
```

Fig. 5. Checking for the presence of empty values

Figure 6 shows the output of the first 5 rows of the input data table.



```
   USMER  MEDICAL_UNIT  SEX  PATIENT_TYPE  PNEUMONIA  AGE  DIABETES  COPD  \
0      0             1    1             0          1   40         0     0
1      0             1    1             1          0   64         0     0
2      0             1    1             1          1   64         1     0
3      0             1    1             0          0   37         1     0
4      0             1    1             0          0   25         0     0

   ASTHMA  INMSUPR  HIPERTENSION  OTHER_DISEASE  CARDIOVASCULAR  OBESITY  \
0       0        0             0              0               0        0
1       0        0             0              0               0        0
2       0        1             1              0               0        0
3       0        0             1              0               0        1
4       0        0             0              0               0        0

   RENAL_CHRONIC  TOBACCO  CLASIFFICATION_FINAL
0              0        0                     1
1              0        0                     1
2              1        0                     1
3              0        0                     1
4              0        0                     1
```

Fig. 6. View of the first 5 rows of input medical data

3.3. Feature selection

We should note that the Shannon method estimates the informativeness of the investigated recognition in a normalized quantity, which takes values from 0 to 1. Comparison of results of both methods allows the following conclusions: the considered methods do not contradict each other and give similar sets of the most informative features on the same training samples, and the results of the Shannon and Kullback methods mostly coincide. Table shows the results of using methods for assessing the informativeness of features.

Results of calculating the informativeness of features

Name	Results (Shannon)	Results (Kullback–Leibler)
Treatment in medical institutions	0.92	1.55
Medical insurance	0.44	1.99
Gender	0.99	1.73
Patient type	0.55	1.97
Pneumonia	0.43	0.94
Age	0.86	2.00
Diabetes	0.46	0.99
Chronic lung disease	0.08	0.00
Asthma	0.19	0.46
Weakness of the immune system	0.09	0.025
High blood pressure	0.57	1.13
Another disease	0.16	0.34
Cardiovascular disease	0.12	0.18
Obesity	0.60	1.17
Chronic kidney disease	0.10	0.08
Smoking	0.40	0.89
Covid-19 disease	0.93	1.56

The obtained results were visualized. Figures 7 and 8 show which features have an impact and informativeness and which can be excluded from the set.

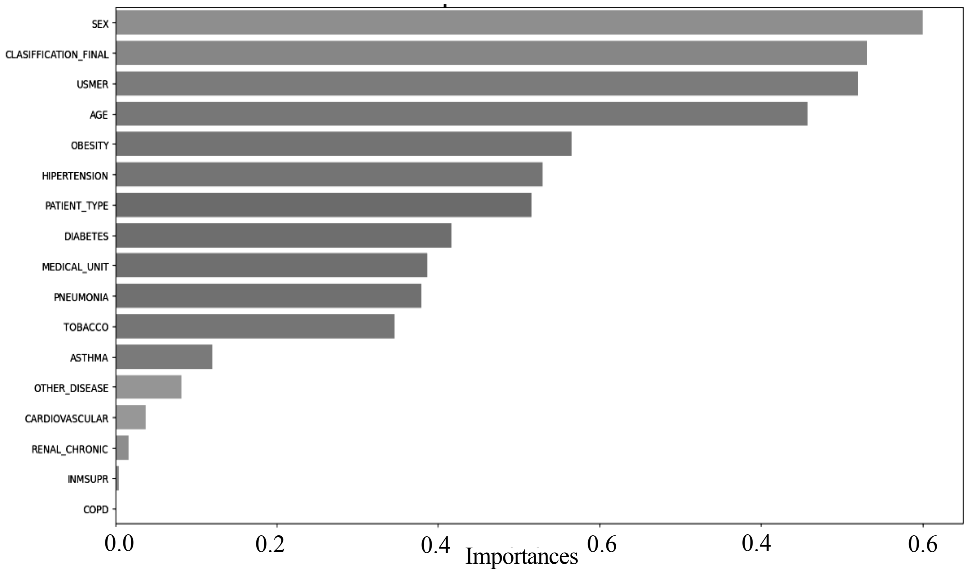


Fig. 7. Diagram of informativeness assessment by the Shannon method

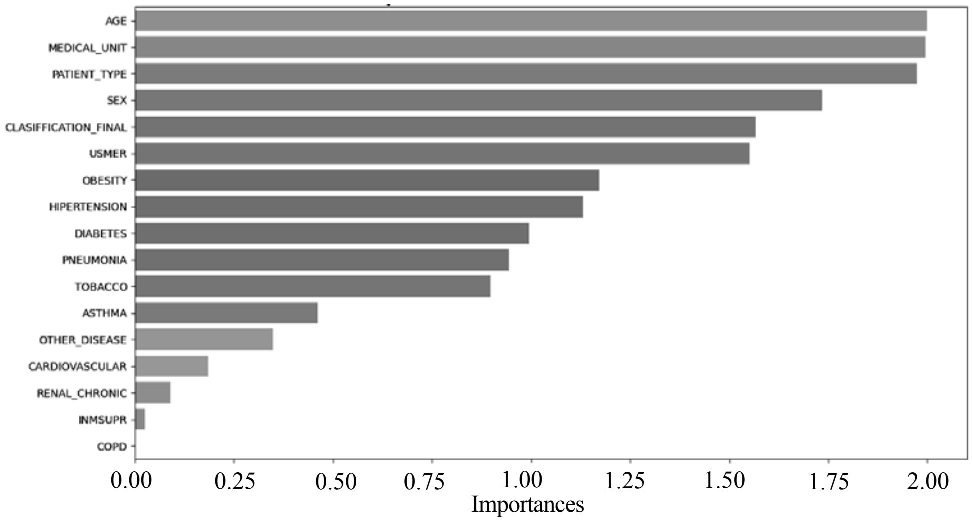


Fig. 8. Diagram of informativeness assessment by the Kullback–Leibler method

4. DISCUSSION

The evaluation of informativeness is pivotal in understanding the dynamics of epidemic processes and devising effective disease control strategies. This study aimed to implement and evaluate methods to assess the informativeness of features that influence epidemic processes. The methods examined in this study, namely the Shannon method and the Kullback–Leibler method, are grounded in the principles of information theory and have distinct advantages, differences, and commonalities. Both methods utilize the concept of event probability and employ a logarithmic scale to measure informativeness, which is particularly helpful when dealing with extremely small or large probability values. These methods are also extensively applied in machine learning for feature selection, model management, and assessing feature informativeness.

The study found that the Shannon and Kullback–Leibler methods are valuable tools for quantifying the information contained in a random process and thus can be applied across various fields such as information theory, statistics, and machine learning. The comparison of different methods and the results they yield is crucial for understanding their applicability and limitations. It was observed that certain features, such as "Chronic lung disease," "Chronic kidney disease," and "Weakness of the immune system," did not carry significant information for further analysis and prediction, indicating that not all available features are necessarily informative or relevant for epidemic process analysis.

Developing an information system that facilitates the assessment of feature informativeness is a significant contribution of this study. This system not only supports data sample uploading but also enables the calculation of the informativeness of factors that influence the epidemic process. The visualization of the system's results aids in the interpretation and application of the findings.

However, there are several limitations to this study. First, the analysis was based on a specific data set, and the informativeness of features may vary in different contexts or with different diseases. Therefore, the findings of this study may not be directly generalizable to other epidemic processes. Second, the study focused on two specific methods of assessing informativeness, and there may be other methods that could yield different results or insights. Additionally, the study did not consider the potential interactions between different features, which could also influence the informativeness of individual features.

The study contributes a novel perspective by demonstrating a methodical approach to assess the informativeness of various features related to epidemic processes. By applying the Shannon and Kullback–Leibler methods, this study brings a quantitative, data-driven approach to a field often dominated by qualitative assessments and heuristic methods. This quantitative approach can lead to more objective, replicable, and actionable insights into the drivers of epidemic processes.

Additionally, this study contributes by identifying specific features that are not informative in the context of the analyzed data set. This is crucial as it challenges conventional wisdom and prompts a re-evaluation of commonly held beliefs about the most critical factors in driving epidemic processes. This can lead to a paradigm shift in how epidemic processes are analyzed and managed, moving away from a one-size-fits-all approach to a more nuanced, data-driven approach.

Moreover, the study compares two widely used methods for assessing informativeness, thereby providing insights into their relative merits and limitations. This can guide researchers and practitioners in selecting the most appropriate method for their specific context and research questions.

Developing an information system that supports data upload and informativeness calculations adds a practical tool that researchers and practitioners can use to assess the informativeness of features in their own data sets. This contributes to the methodological rigor of future studies and enhances the practical applicability of the findings by enabling real-world implementation.

Future research should validate the findings of this study in different contexts and with different diseases to assess the generalizability of the results. It would also be beneficial to compare the performance of the Shannon and Kullback–Leibler methods with other methods of assessing informativeness. Furthermore, future studies should also explore the potential interactions between different features and their impact on the informativeness of individual features.

Developing and evaluating more sophisticated information systems that can account for feature interactions and other complexities in the data would be a valuable avenue for future research.

Overall, this study contributes a novel perspective, challenges conventional wisdom, provides practical insights into the relative merits of different methods, and offers a practical tool for assessing feature informativeness. These contributions are crucial for enhancing our understanding of epidemic processes and developing more effective strategies for their management.

CONCLUSIONS

The use of methods for assessing informativeness is crucial in analyzing epidemic processes. The main objective of such an analysis is to understand the spread of the disease and determine the effectiveness of strategies to combat it. Methods of informativeness assessment allow for determining how well a specific parameter correlates with the risk of disease. This enables identifying population groups that may be more susceptible to the disease and considering this when developing prevention and treatment strategies.

As a result of this study, methods were identified and implemented that allow assessing the informativeness of features. Methods for assessing the informativeness of features were considered; algorithmic models were developed for the Kullback–Leibler and Shannon methods. Both considered methods are based on information theory principles and have advantages, differences, and standard features. Thus, both the Shannon method and the Kullback–Leibler method are based on the concept of the probability of events, use a logarithmic scale to measure informativeness, which helps in dealing with very small or tremendous probability values, and is widely used in the field of machine learning for evaluating the informativeness of features, model management, and feature selection. Overall, the Shannon and Kullback–Leibler informativeness assessment methods are valuable tools for measuring the information contained in a random process. They can be used in various fields, such as information theory, statistics, machine learning, etc.

Specific examples of using the described algorithmic models are presented. A comparison of different methods and their results was carried out. It was found that such features as “Chronic lung disease”, “Chronic kidney disease”, and “Weakness of the immune system” do not carry information for further work with the table and burden the prediction relative to the presented data set.

An information system for analyzing epidemic process data was developed to assess the informativeness of features. This system supports data sample uploading and calculations of the informativeness of factors affecting the epidemic process. The results of the system operation are visualized.

Acknowledgements. The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic “Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management”.

REFERENCES

1. K. Batko and A. Ślęzak, “The use of Big Data Analytics in healthcare,” *Big Data*, vol. 9, no. 1 (2022), https://doi.org/10.1186/s40537-021-00553-4.
2. I. Izonin, R. Tkachenko, I. Dronyuk, et al., “Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method,” *Mathematical Biosciences and Engineering*, vol. 18, no. 3, pp. 2599–2613 (2021), https://doi.org/10.3934/mbe.2021132.

3. S.Y. Lee, B. Lei, and B. Mallick, "Estimation of COVID-19 spread curves integrating global data and borrowing information," *PLOS ONE*, vol. 15, no. 7, 0236860 (2020), https://doi.org/10.1371/journal.pone.0236860.
4. S. Ma, Y. Sun, and S. Yang, "Using Internet Search Data to Forecast COVID-19 Trends: A Systematic Review," *Analytics*, vol. 1, no. 2, pp. 210–227 (2022), https://doi.org/10.3390/analytics1020014.
5. A. Ibrahim, U. W. Humphries, A. Khan, et al., "COVID-19 Model with High- and Low-Risk Susceptible Population Incorporating the Effect of Vaccines," *Vaccines*, vol. 11, no. 1 (2022), https://doi.org/10.3390/vaccines11010003.
6. N. Davidich, I. Chumachenko, Y. Davidich, et al., "Advanced Traveller Information Systems to Optimizing Freight Driver Route Selection," *2020 13th International Conference on Developments in eSystems Engineering (DeSE)* (2020), https://doi.org/10.1109/dese51703.2020.9450763.
7. S. Fedushko and T. Ustyianovych, "E-Commerce Customers Behavior Research Using Cohort Analysis: A Case Study of COVID-19," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 8, no. 1, pp. 1-12 (2022), https://doi.org/10.3390/joitmc8010012.
8. P.S. Knopov, O.S. Samosonok, and G.D. Bila, "A Model of Infectious Disease Spread with Hidden Carriers," *Cybernetics and Systems Analysis*, vol. 57, no. 4, pp. 647–655 (2021), https://doi.org/10.1007/s10559-021-00390-6.
9. D.A. Klyushin, "Effective algorithms for solving statistical problems posed by COVID-19 pandemic," *Elsevier eBooks*, pp. 21–44 (2023), https://doi.org/10.1016/b978-0-323-90531-2.00005-9.
10. I. Krak, H. Kudin, V. Kasianiuk, et al., "Hyperplane Clustering of the Data in the Vector Space of Features Based on Pseudo Inversion Tools," *CEUR Workshop Proceesings*, vol. 3003, pp. 98–105 (2021), https://ceur-ws.org/Vol-3003/short4.pdf
11. O. Filchakova, D. Dossym, A. Ilyas, et al., "Review of COVID-19 testing and diagnostic methods," *Talanta*, vol. 244, 123409 (2022), https://doi.org/10.1016/ j.talanta.2022.123409.
12. S. Patil, H. Lu, C. L. Saunders, et al., "Public preferences for electronic health data storage, access, and sharing — evidence from a pan-European survey," *Journal of the American Medical Informatics Association*, vol. 23, no. 6, pp. 1096–1106 (2016), https://doi.org/10.1093/jamia/ocw012.
13. V. Berisha, C. Krantsevich, P. R. Hahn, et al., "Digital medicine and the curse of dimensionality," *npj Digital Medicine*, vol. 4, no. 1 (2021) https://doi.org/10.1038/ s41746-021-00521-5.
14. K. Bazilevych, S. Krivtsov, and M. Butkevych, "Intelligent Evaluation of the Informative Features of Cardiac Studies Diagnostic Data using Shannon Method," *CEUR Workshop Proceedings*, vol. 3003, pp. 65–75 (2021).
15. I. Meniailov and H. Padalko, "Application of Multidimensional Scaling Model for Hepatitis C Data Dimensionality Reduction," *CEUR Workshop Proceedings*, vol. 3348, pp. 34–43 (2022).
16. K. O. Bazilevych, D. I. Chumachenko, L. F. Hulianytskyi, et al., "Intelligent Decision-Support System for Epidemiological Diagnostics. I. A Concept of Architecture Design," *Cybernetics and Systems Analysis*, vol. 58, no. 3, pp. 343–353 (2022), https://doi.org/10.1007/s10559-022-00466-x.
17. K.O. Bazilevych, D.I. Chumachenko, L.F. Hulianytskyi, et al., Intelligent Decision-Support System for Epidemiological Diagnostics. II. Information Technologies Development," *Cybernetics and Systems Analysis*, vol. 58, no. 4, pp. 499–509 (2022). https://doi.org/10.1007/s10559-022-00484-9
18. D. Panda, R. Ray, and Satya Ranjan Dash, "Feature Selection: Role in Designing Smart Healthcare Models," *Intelligent systems reference library*, vol. 178, pp. 143–162, (2020), https://doi.org/10.1007/978-3-030-37551-5_9.
19. D. Geiszler, D. A. Polasky, F. Yu, and A. I. Nesvizhskii, "Detecting diagnostic features in MS/MS spectra of post-translationally modified peptides," *Nature Communications*, vol. 14, no. 1 (2023), https://doi.org/10.1038/s41467-023-39828-0.
20. D.E. Ehrmann, S. Joshi, S.D. Goodfellow, et al., "Making machine learning matter to clinicians: model actionability in medical decision-making," *npj Digital Medicine*, vol. 6, no. 1 (2023), https://doi.org/10.1038/s41746-023-00753-7.
21. O. Cliff, M. Prokopenko, and R. Fitch, "Minimising the Kullback–Leibler Divergence for Model Selection in Distributed Nonlinear Systems," *Entropy*, vol. 20, no. 2, p. 51 (2018), doi: https://doi.org/10.3390/e20020051.

22. X. Wang, W. Hou, H. Zhang, et al., “KDE-OCSVM model using Kullback–Leibler divergence to detect anomalies in medical claims,” *Expert Systems with Applications*, vol. 200, 117056 (2022), doi: https://doi.org/10.1016/j.eswa.2022.117056.
23. N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, et al., “A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction,” *Frontiers in Bioinformatics*, vol. 2 (2022), https://doi.org/10.3389/fbinf.2022.927312.
24. J. Li, K. Cheng, S. Wang, et al., “Feature Selection,” *ACM Computing Surveys*, vol. 50, no.6, pp. 1–45 (2018), https://doi.org/10.1145/3136625.
25. F. Jalali-najafabadi, M. Stadler, N. Dand, et al., “Application of information theoretic feature selection and machine learning methods for the development of genetic risk prediction models,” *Scientific Reports*, vol. 11, no. 1 (2021), https://doi.org/10.1038/s41598-021-00854-x.
26. A. D. Al-Nasser, A. Rawashdeh, and A. Talal, “On using Shannon entropy measure for formulating new weighted exponential distribution,” *Journal of Taibah University for Science*, vol. 16, no. 1, pp. 1035–1047 (2022), https://doi.org/10.1080/ 16583655.2022.2135806.
27. “*Scikit-learn: machine learning in Python,*” *Scikit-learn.org* (2019), https://scikit-learn.org/stable/
28. “COVID-19 Dataset,” *www.kaggle.com* (2022), https://www.kaggle.com/datasets/ meirnizri/covid19-dataset

Received 06.09.2023

INFORMATION ON THE ARTICLE

Kseniia O. Bazilevych, ORCID: 0000-0001-5332-9545, National Aerospace University “Kharkiv Aviation Institute”, Ukraine, e-mail: k.bazilevych@khai.edu

Olena Yu. Kyrylenko, ORCID: 0009-0005-8917-0878, National Aerospace University “Kharkiv Aviation Institute”, Ukraine, e-mail: o.kyrylenko@khai.edu

Yurii L. Parfenyuk, ORCID: 0000-0001-5357-1868, V.N. Karazin Kharkiv National University, Ukraine, e-mail: parfuriy.l@gmail.com

Sergiy V. Yakovlev, ORCID: 0000-0003-1707-843X, National Aerospace University “Kharkiv Aviation Institute”, Ukraine, e-mail: s.yakovlev@khai.edu

Serhii O. Krivtsov, ORCID: 0000-0001-5214-0927, National Aerospace University “Kharkiv Aviation Institute”, Ukraine, e-mail: krivtsovpro@gmail.com

Ievgen S. Meniailov, ORCID: 0000-0002-9440-8378, V.N. Karazin Kharkiv National University, Ukraine, e-mail: evgenii.menyailov@gmail.com

Victoriya O. Kuznietcova, ORCID: 0000-0003-3882-1333, V.N. Karazin Kharkiv National University, Ukraine, e-mail: vkuznietcova@karazin.ua

Dmytro I. Chumachenko, ORCID: 0000-0003-2623-3294, National Aerospace University “Kharkiv Aviation Institute”, Ukraine, e-mail: d.chumachenko@khai.edu

ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ ОЦІНЮВАННЯ ІНФОРМАТИВНОСТІ ОЗНАК ЕПІДЕМІЧНОГО ПРОЦЕСУ / К.О. Базілевич, О.Ю. Кіріленко, Ю.Л. Парфенюк, С.В. Яковлев, С.О. Кривцов, Є.С. Меняйлов, В.О. Кузнецова, Д.І. Чумаченко

Анотація. Роботп полягає в оцінюванні інформативності параметрів, які впливають на епідемічні процеси, з використанням методів Шенона та Кульбака–Лейблера на основі їх фундаментальності у принципах теорії інформації та їх широкого застосування в машинному навчанні, статистиці та інших відповідних галузях. Проведено порівняльний аналіз результатів, отриманих обома методами, розроблено інформаційну систему для спрощення завантаження вибірок даних та обчислення інформативності факторів, які впливають на епідемічні процеси. Показано, що деякі ознаки, такі як «хронічне захворювання легень», «хронічне захворювання нирок» та «ослаблений імунітет», не містили значущої інформації для подальшого аналізу та ускладнювали процес прогнозування за даними досліджуваного набору даних. Розроблена інформаційна система ефективно підтримує оцінювання інформативності ознак, тим самим сприяючи комплексному аналізу епідемічних процесів, візуалізації результатів, а також поточному стану знань. Надано конкретні приклади застосування описаних алгоритмічних моделей, порівняння різних методів та їх результатів та розроблення підтримувального інструменту для аналізу епідемічних процесів.

Ключові слова: інформаційна система, епідемічний процес, інформативність ознаки, метод Шенона, метод Кульбака–Лейблера.

ЕВРИСТИЧНІ МЕТОДИ ТА АЛГОРИТМИ В СИСТЕМНОМУ АНАЛІЗІ ТА УПРАВЛІННІ

UDC 62-50
DOI: 10.20535/SRIT.2308-8893.2023.4.09

SURVEY OF IMAGE DEDUPLICATION FOR CLOUD STORAGE

S. CHAUDHARI, R. APARNA

Abstract. Increased growth of real-life communication has motivated the creation, transmission, and digital storage of vast volumes of images and video data on the cloud. The explosive increase in virtual/visual image data on cloud servers requires efficient storage utilization that can be addressed using image deduplication technology. Even though the virtual and visual image properties are different, the existing literature uses a similar approach for deduplication checks, which motivated us to consider both image types for this review. This article aims to provide a detailed survey of state-of-the-art visuals as well as virtual image deduplication techniques in a cloud environment, summarizing and organizing them by developing a five-dimensional taxonomy for analysing the features and performance with several non-overlapping categories in each dimension. These include: 1) location of applying deduplication; 2) image feature extraction; 3) time of application; 4) image data partitioning strategy; 5) involvement of user dataset level. Existing image deduplication techniques are categorized into two main categories based on whether the technique involves security. A comparison of techniques is discussed across a set of functional and performance parameters. The current issues are highlighted with the possible future directions to motivate further research studies on the topic.

Keywords: image deduplication, cloud computing, cloud storage, image copy detection.

INTRODUCTION

With the massive development of electronics and the internet, digital data is increasing at an alarming rate. This includes data in the form of text, images, videos, sketches, etc. All this data comes from different parts of the Internet and hence causes information explosion due to huge velocity of data generation and huge variety of data sources. In 2007, it is said that the total digital resources of the world exceeded the global storage capacity for the very first time. Hence, it was decided that this problem of information explosion cannot be handled by simply increasing the amount of storage. But now it is estimated that by 2025, there will be 163.2 zettabytes of digital data [35].

Primary data generators like social networking platforms, industries and transactional data from various businesses are generating huge volumes of data every day. Due to the sudden increase in volumes of data, it becomes extremely crucial to be able to store this data in a cost-effective manner that optimizes storage. Cloud based infrastructure for on demand service provisioning from anywhere, anytime is the popular solution used. The National Institute of Standards and Technology (NIST) reference architecture for cloud computing has following five actors: 1) cloud consumer; 2) cloud provider; 3) cloud carrier; 4) cloud auditor; 5) cloud broker. The interaction among these actors is shown in Fig. 1 along with their activities and functions.

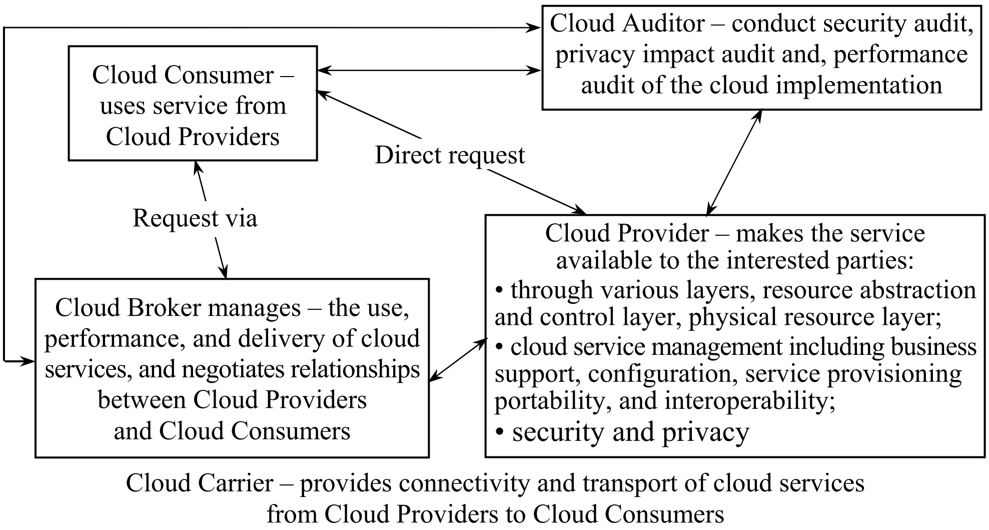

Cloud Auditor – conduct security audit, privacy impact audit and, performance audit of the cloud implementation
Cloud Consumer – uses service from Cloud Providers
Direct request
Request via
Cloud Provider – makes the service available to the interested parties:
• through various layers, resource abstraction and control layer, physical resource layer;
• cloud service management including business support, configuration, service provisioning portability, and interoperability;
• security and privacy
Cloud Broker manages – the use, performance, and delivery of cloud services, and negotiates relationships between Cloud Providers and Cloud Consumers
Cloud Carrier – provides connectivity and transport of cloud services from Cloud Providers to Cloud Consumers

Fig. 1. Cloud Actors Activities and Interaction in Cloud

Storage as a service on Cloud is one of the critical and popular services wherein the cloud storage provider provides cost effective and easy to access storage space on the cloud to the interested customers to host their data instead of maintaining it on their on-premises. The user data stored on the cloud can be in any form like images, audio, video etc. The customers or the data owners cannot rely on public service providers like cloud for data security. So, to provide security to the data, many researchers proposed secure storage techniques for storing the data on cloud.

The cloud services are provided through virtual images whose size is very large requiring large amount of storage space in addition to huge network transmission requirements and reduction in operation time. Virtual images, each with large size starting with 1GB with different configurations may belong to a single cloud user. Almost 80% of the virtual image content is identical among these Virtual Machine (VM) images due to existence of similar data segments [1]. The two primary reasons namely sudden explosion of data and similarity in virtual images apparently induce a need in Cloud Service Providers (CSP) to optimize the storage and network bandwidth used in data transfer of VM images.

Hence, a concept called deduplication was formulated, which could identify duplicates and delete all copies except one (or precisely retain as many copies as specified by deduplication ratio) [36]. It optimally minimizes the storage utilization by deleting redundant data from the cloud storage or data centers and thereby bringing down the unnecessary usage of network bandwidth [40]. It is a lossless data compression technology, which replaces duplicate image copies by using pointers to the unique image object.

The image deduplication [40] process involves four steps to remove duplicate images or parts of images as follows [39]: 1) file chunking wherein the image is divided into fixed or variable size blocks known as chunks; 2) fingerprint generation: the fingerprint will be computed using some transformation algorithm/technique such as hash function; 3) fingerprint lookup: the fingerprint of already existing images will be compared with this newly created fingerprint in step 2 for identifying the duplicate file/block. If it is found to be same, this new file/block will be discarded otherwise it will be stored in the cloud storage; 4) data storage: store the unique image/block systematically on the cloud storage.

Deduplication can be done at two levels: the single user level and the cross level. In the single user level, the deduplication is done keeping only one user in mind and duplication happens in their own storage. Cross level deduplication is when data is compared by taking from many users, and then redundant data is deleted. It can also be done at either the client side or the server side. At the client side, the data is checked for duplicates by the client itself and is then sent to the server. At the server side, the server collects all the information from the client and then duplicates are found and removed. Deduplication also has two feature-based methods known as global feature-based method and local feature-based method.

There is no detailed investigation done till now to review image data deduplication techniques and its characteristics. Few non-standard articles exist explaining data deduplication survey in unstructured way. The authors of [37] have classified the existing data deduplication techniques into two categories as source deduplication and target deduplication. Source deduplication is further classified into file-based and sub-file-based. Sub file is further considered as fixed or variable length. Target deduplication is further classified as post-process and inline. They have discussed another way for classifying data deduplication namely offline and online deduplication. Even though many research articles exist, the paper discusses about only 10 research articles on data deduplication. The authors of [38] discuss 14 deduplication approaches without any taxonomy or relation among them. They have included 24 research articles under these approaches and compare them in terms of scalability, throughput, efficiency, amount of used bandwidth and cost. Many comparison parameters could have been considered along with some more deduplication techniques. Lack of systematic review/survey motivated us to do this work.

In this survey paper, we survey different types of deduplications or copy detection that have been done for images in a systematic and structured way. As important as it is, traditional deduplication mechanisms can only be used if two images have the same bit stream, that is it can only be used if two images are completely the same. It does not apply for an image that has been cropped, rotated, or edited out. Automatic methods are now getting more attention with the increase in the redundant information. Also, cloud computing has proved to be very flexible and economical service provider that provide to maintain huge amount of data. In this world of immense data, users normally upload similar images in different storages either due to storage restrictions or network restrictions. The aim of this paper is to understand and observe the different techniques used for image deduplication in terms of functional and performance parameters.

Our contributions. Consequently, this significant amount of published research on Image deduplication requires some categorization to provide convenient overview of the current state of the art. To this end, we have developed multi-dimensional taxonomy to classify the Image deduplication research based on the properties supported in the research work as described in Section 2. Even though multi-dimensional taxonomy is used popularly for defining image deduplication techniques, we categorize them into two main categories based on whether security is incorporated or not. Non-secure techniques are further categorized based on image type as virtual image or pixel image. Techniques in each category is discussed across a set of functional and performance parameters. The presented taxonomy allows us to analyze the Image deduplication research trends over time and various features supported in the work. To illustrate the usefulness of the provided classification, we discuss a detailed survey of the collected research articles from extensive databases available online where Image deduplication-based references can be explored according to the designed dimensions and categories of the presented taxonomy.

Our specific contributions are as follows: 1) design and discuss multi-dimensional taxonomies for comparison of the various parameters used in image deduplication; 2) explain the image deduplication research trends across two main categories based on whether security is considered or not. Non-secure techniques are further categorized based on image type as virtual image or pixel image; 3) compare the discussed image deduplication schemes in each category in terms of functional and performance parameters.

The remaining part of this article is structured in various sections as follows. Section 2 explains the methodology for creating the taxonomy of Image deduplication research work with its dimensions and categories. It also explains the Image deduplication-related research articles to analyze and provide trends on the distribution across the proposed dimensions. Section 3 presents a detailed survey of the key research findings and related comparison with respect to a set of functional and performance parameters related to Image deduplication. Section 4 addresses the scope of the research on Image deduplication. Finally, conclusions are drawn in Section 5.

DESIGN OF IMAGE DEDUPLICATION TAXONOMY AND CLASSIFICATION

The taxonomy is aimed at classifying the work carried out in Image deduplication to have an in-depth understanding of the topic. Taxonomy construction varies from topic to topic, but all works in one class given in the taxonomy should be similar in the features or properties. The classification categories should be non-overlapping with well-defined limits between them. The taxonomy designed for Image deduplication related research for analyzing the features and performance includes five dimensions with several non-overlapping categories in each dimension. These include: 1) location of applying deduplication; 2) image feature extraction; 3) time of application; 4) image data partitioning strategy; 5) involvement of user dataset level.

Each dimension consists of a set of categories used to classify the existing image deduplication related articles. The presented taxonomy allows us to analyze the image deduplication research trends over time and various features supported in the work. A given article may not be mutually exclusive to the category as it may belong to one or more categories per dimension. The illustration of image deduplication taxonomy in graphical form is shown in Fig. 2. We have tried to minimize the possible overlap between the existing image deduplication techniques as per the proposed dimensions in this early stage of defining the classification categories.

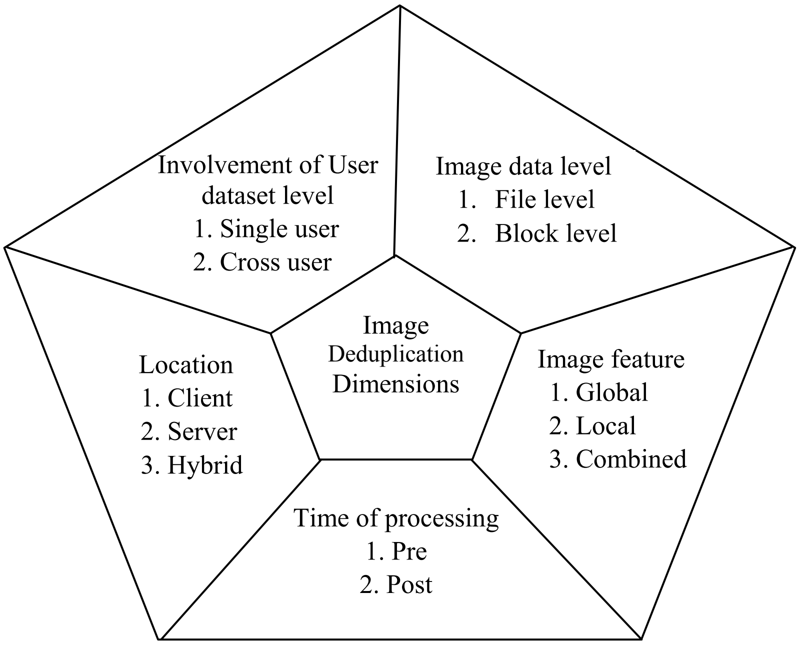


Fig. 2. Taxonomy of Image Deduplication Techniques

The first dimension namely location of deduplication in proposed image deduplication taxonomy further classifies the existing research works into three categories depending upon the place where deduplication is carried out. Since all the papers are based on client server architecture, cloud being the serving platform, the process of deduplication can be executed at the client side, server side or partly in both the places. We categorize the image deduplication techniques with respect to location of deduplication dimension into three classes as explained next: 1) server-side image deduplication: users upload the images to the cloud server in server-side image deduplication category and then the cloud service provider will perform the image deduplication check on its cloud storage to identify whether newly arrived image already exists on the server or not; 2) client-side image deduplication: client will identify the existence of similar data on the cloud server before sending it entirely to the cloud in client-side image deduplication category. Server-side image deduplication reduces computational cost at the client side but with high bandwidth requirements; 3) hybrid location-based image deduplication: Image deduplication check may be partially done at client side whereas remaining check will be done at the server side in hybrid image deduplication category of this dimension.

Second dimension named as image feature based, as proposed in image deduplication taxonomy further classifies the existing research works into three categories depending upon the usage of local and global features of images for identification of deduplication. Image features are numerical values extracted from images that are used as discriminating information to distinguish various images or parts of images. Features are extracted for reducing the processing overhead as they are small when compared to image data. Global features of image describe the whole image to generalize the image data while local image features describe small group of pixels in image. Combination of both improves the accuracy of image recognition with the side effect of computational overhead. We categorize the image deduplication techniques in image feature-based dimensions into three classes namely: 1) global feature-based image deduplication: global features of an image are used to identify the image deduplication; 2) local feature-based image deduplication: local features of an image are used to identify the image deduplication; 3) combined feature-based image deduplication: local and global features of an image are used to identify the image deduplication.

As per the proposed image deduplication taxonomy, third dimension identified as Time of duplicate removal processing, further classifies the existing research works into two categories depending upon the time at which the duplicate data is removed for identified deduplicated images, as explained next: 1) inline image deduplication processing: the identification of deduplication is immediately started when cloud server receives the image without storing it. The deduplicated image/block of image is deleted before storing for achieving unique image data copy; 2) post-image deduplication processing: the received image will be stored in buffer on the cloud server first, then the deduplication check will be performed to identify the duplicate image/block of image. Only the unique images/blocks will be stored on the cloud server database/storage.

This dimension namely Time of duplicate removal with respect to virtual image deduplication can be categorized into three categories namely deduplication before backup, deduplication during backup and deduplication after backup. In the first case namely deduplication before backup, duplicate check is done before performing the backup operation so that the size of the data transmitted would be that of the compressed image size. Here, both the fingerprint calculation and index lookup operation must be performed by the host node. In the second case namely deduplication after backup, deduplication check is performed after backing up the image. Since whole image is transmitted, the data transmission size would be large. In this case, storage node is the location for the fingerprint calculation and index lookup operation. The third case namely deduplication operation during backup aims at balancing the resource overhead at both the host side and storage side.

Fourth dimension named as Image data level in the proposed image deduplication taxonomy further classifies the existing research works into three categories depending upon the whole image or part of image being used for identification of duplicates. The categories in this dimension are given as follows: 1) file-level image deduplication: the same image existing on the cloud server will be checked using the hash value created for each file based on the specific hash function. If the received image hash value and one of the existing image hash values is same, then the received image will not be stored otherwise it will be stored on cloud server database; 2) block level image deduplication: the received image will be divided into blocks. Hash value is calculated for each block using specific hash function. The hash value for the block is called as block fingerprint. Only one block will be stored on cloud server for two or more blocks with same fingerprint. Otherwise, all blocks are stored on the cloud server; 3) hybrid-level image deduplication: both file – level and block level hash are checked for image deduplication check.

Fifth dimension named as involvement of user dataset level in proposed image deduplication taxonomy further classifies the existing research works into two categories depending upon the usage of user databases being scanned for checking identical images. Dataset used for checking image deduplication may belong to specific user or may have permission to store data of multiple users. We categorize the image deduplication techniques based on involvement of user dataset level dimensions into two classes namely: 1) single user level image deduplication: image dataset belonging to a user is scanned to check the duplicate images for that user alone; 2) cross-user level image deduplication: Image databases of multiple users are scanned to check the duplicate image. Even though cross user level image deduplication generates higher deduplication ratio and is attractive in terms of storage cost in comparison with single user level image deduplication, it affects the privacy and security concern for users.

Even though multi-dimensional taxonomy is used popularly for defining image deduplication techniques in the literature in a scattered way, we categorize them into two main categories based on whether security is incorporated or not. Non-secure techniques are further categorized based on image type as virtual image or pixel image as shown in Fig. 3.

LITERATURE SURVEY ON IMAGE DEDUPLICATION TECHNIQUES

This section discusses the two main categories designed in our taxonomy as shown in Fig.3 based on whether the techniques incorporate security or not. Non-secure approaches are further categorized for virtual image or pixel image types. Non-secure image deduplication techniques are described in Section 3.1 and Section 3.2 for virtual image types and pixel types respectively while Section 3.3 discusses all secure techniques.

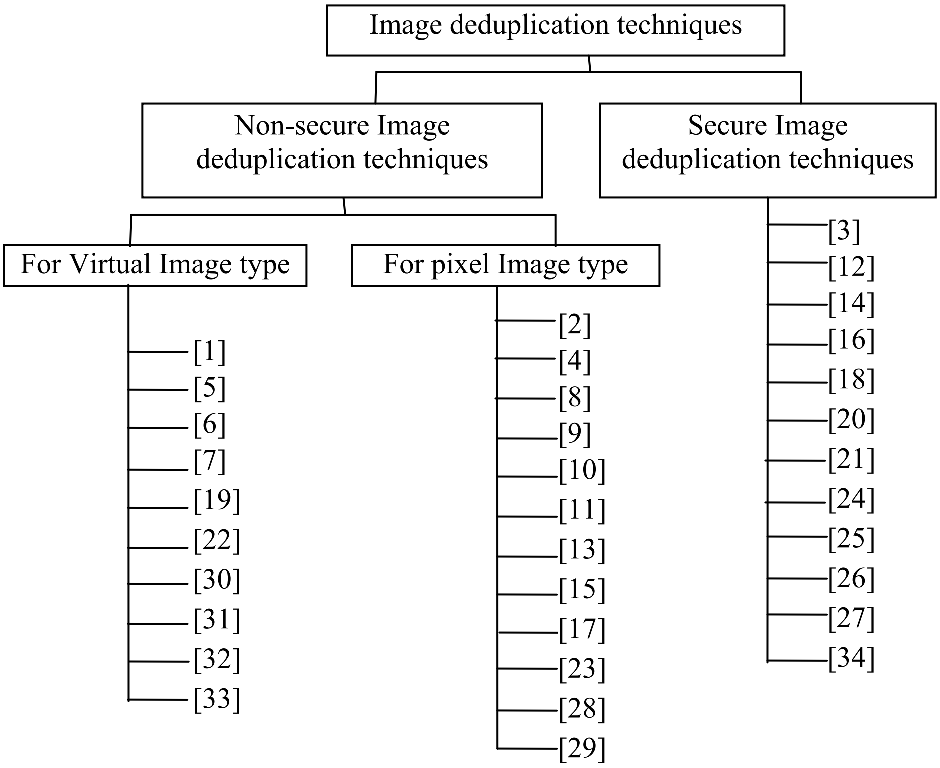


Fig. 3. Proposed Taxonomy for Image Deduplication Techniques

As the proposed image deduplication techniques are not mutually exclusive to any specific dimension, one technique may belong to one or more dimensions.

They are compared in terms of functionality based on the dimensions and performance parameters in the respective Section. Table 1 provides the functionality based on the dimensions and performance parameters used for comparison of the image deduplication techniques.

NON-SECURE IMAGE DEDUPLICATION FOR VIRTUAL IMAGE TYPE

Virtualization is a very important part of cloud computing. It allows multiple servers running on a single host and all disk contents are encapsulated in a single Virtual Machine (VM) Image. But this mechanism can store redundant data and cause storage issues. A lightweight virtual machine image deduplication backup approach in cloud environment is a technique used to eliminate this problem [1]. The process includes dividing the VM image into chunks and checking if the chunk has a fingerprint. The fingerprint is compared with the existing fingerprints in the fingerprint index table and if it exists then it is not entered otherwise the chunk is added in the storage system. The two problems faced are as follows: 1) if the fingerprint index table is long then it will take longer to compare the fingerprints; 2) the process can interrupt the other processes as different virtual machines take the same runtime. This paper gives a classification method to reduce the fingerprint search by converting global duplication to local duplication and to improve index lookup. Two sampling methods are used to find the proper group to perform the deduplication operation in the virtual machine image. A numerical method is also used to calculate the ample space size. Deduplication rate is 10.2%.

Table 1. Comparison parameters for image deduplication techniques

SN	Functional parameter	Remark
1	Matching Algorithm Used	Scale invariant feature transform (SIFT), principal component analysis (PCA) for SIFT, min-hash algorithm, feature extraction, high dimension indexing, accuracy optimization, centroid selection, deduplication evaluator, mean median, standard deviation, hash, map reduce, CRC, pixel based, special layout, visual similarity
2	Location	Client side, server side, hybrid
3	Processing time	Post-process, inline
4	Feature-based	Local, global, structural features, visual model features, and feature points
5	Image data level	File, Block,Hybrid
6	Block size	Fixed length, variable length
7	Image content type	Pixel image (PI) or Virtual Machine Image (VMI)
8	Metadata overhead	The extra information required to be stored along with actual image data
9	Optimization objective	The goal of the proposed image deduplication technique
Cloud environment parameter		
10	Cloud/OS - Cloud software	If any specific cloud software used
11	Cloud type	Public, private and hybrid cloud
12	Number of cloud images	Finite number of images existing in the cloud database
13	User level	Single user or cross user
14	Security provisioning Protocol	Usage of cryptographic protocol, adaptation of cryptographic protocol

An improved k-means clustering method is implemented in Clustering-based acceleration for virtual machine image deduplication in the cloud environment [5]. This is the first paper to have image layout taken into consideration and propose the method of small group merging and periodical triggering to store the virtual machine deduplication. Experimental results show the robustness and efficiency of this method. Deduplication rate is 89.74%.

The number of virtual machine and images grows very rapidly and takes a lot of storage space out of which 90% of the data is redundant. The storage problems caused is studied an Improved Image File Storage Method Using Data Deduplication [6] and discussions about employing deduplication and evaluation. A reference count for image is added to show reliability of image libraries.

IM-dedup proposed in [7] transmits the unique blocks of image to the cloud server for reducing transmission time. The kernel file system with deduplication functionality in the image storage helps to manage the duplicated blocks through indexing. Client and server communicate within each other during the process of image deduplication.

Deduplication-Enabled P2P based VM Image distribution protocol is introduced to speed up the provisioning in the VM [19]. Peer 1 contacts a tracker which sends it the list of its peers which also has an image of file A, it also shows the similarity Matrix between two peers.

Scalable read/write throughput in RAID with deduplication capability to Ext4 file system is proposed in [22] as deduplication file system named as ScaleDFS. Parallel processing for fingerprints computation on multiple CPU core improves the write throughput. Deduplication cache improves read throughput and retrieve identical blocks easily. Reduced memory usage cache more fingerprint information in memory. Deduplication is focused for single storage partition file system.

Authors of [30] have proposed an adaptive deduplication mechanism, which performs fixed length and variable length block-level deduplication for reducing VM disk image file size is used in variable length block-level deduplication is implemented using Rabin–Karp rolling hash algorithm. Multithreading in AKKA framework is used to perform the deduplication and streaming since live migration of VM disk image files is a bulky operation.

QuickDedup [31] algorithm is proposed by authors to perform optimal deduplication of VM disk images by reducing the number of hash computations and comparisons and by storing minimal metadata thereby reducing the overall deduplication time. In this approach, a novel byte comparison scheme to create various categories of blocks so that further the QuickDedup algorithm performs the calculation of hashes and their comparisons within the respective categories only. Hence, hash storage space is minimized and comparing within categories speeds up the deduplication of VM disk images much.

Authors of [32] propose a highly parallel deduplication cluster (HPDV) which optimizes VM images by considering the foreground quality of VM services and the background performance of deduplication for VM images. Generally, chunk-based deduplication process involves four sub processes namely chunking, fingerprinting, fingerprint indexing and data storing. Authors of HPDV have parallelized the chunking, fingerprinting tasks which are compute-intensive and fingerprint indexing, which is I/O-intensive task, using the servers in the cluster. Quality of the foreground VM services is ensured while parallelization is in progress by proposing a resource-aware scheduler (RAS) in this work.

Authors of [33] have done an extensive review of several deduplication strategies and come up with a deduplication algorithm for VM images called DedupCloud. The VM images are divided into blocks and stored sequentially in the preprocessing stage. Then the blocks are categorized based on hashes of blocks derived using SHA-3 hash function.

We compare the above discussed non-secure image deduplication techniques for virtual image type in the form of three table. Table 2 gives the comparison in terms of parameters related to algorithms used and perspectives of various dimensions in the image deduplication technique. The comparison parameters include algorithm used for deduplication check, location where the deduplication takes place, application of deduplication, usage of image features, granularity of image data level, block size for block level granularity, content of the image, metadata overhead and objective function of the proposed technique.

Table 2. Comparison of non-secure Image Deduplication techniques for virtual image type

Paper	Matching Algorithm	Place	Processing time	Feature-based	Data level	Block size	Metadata overhead	Optimization objective
1	Improved *k*-means clustering algorithm with fingerprint	Hybrid	Post	Hybrid	Block	Fixed	Fingerprints of fixed size chunks	In memory deduplication by using clustering and sampling of fingerprints – fingerprint search space optimization
5	Improved *k*-means clustering with statistical indices	Server	Post	Local	Block	4KB	Fingerprints of fixed size chunks	Reduce the fingerprint search space and improve the index lookup performance
6	MD5 index	Server	Post	Local	Hybrid	Fixed 4/8KB	MD5 code of entire image file, size of image file, the number of image blocks, file name, storage address of each image block and storage address of the final block	Storage space reduction of VM images
7	MD5 and SHA-1	Client	Inline	Local	Block	Fixed 4KB	Array of fingerprints and the reference counter	Reduction in VMI storage and transmission time
19	Bloom filter's hash function, Rabin fingerprinting scheme	Hybrid	Post	Global	Block	Variable	File block id, hash, and control messages	Minimize data access or transfer during VMI distribution in data centers
22	a POSIX-compliant, kernel-space driver module	Server	42 VM images of different Linux distribution	Cross	Block	Both fixed and variable	Cryptographic fingerprints of blocks, locality (hash) table that holds the full fingerprints and block numbers of the data blocks that correspond to the most recently accessed fingerprint block	Scalable read/write throughput in RAID to provide increased capacity, reliability, and performance. for storage

Table 2. Comparison of non-secure Image Deduplication techniques for virtual image type

Paper	Matching Algorithm	Place	Processing time	Feature-based	Data level	Block size	Metadata overhead	Optimization objective
30	Rabin–Karp rolling hash algorithm	Server	Inline	Cross	Block	Both fixed and variable	Fingerprints of fixed size chunks and thread related metadata	Reduction in image storage space and total migration time and improvement in deduplication rate
31	SHA-1 Hash based on byte comparisons, categorizes the blocks	Server	Post	Local block metadata	Block	Fixed	block numbers, hashes	Least number of hashes and comparisons, minimum metadata, and fast retrieval of VMs for deployment
32	Parallel fingerprint subindexes	Server	Post	Global	Block	Fixed	Fingerprints of fixed size chunks and thread related metadata	Parallelizing chunking and fingerprinting tasks with multiple threads to speed up the tasks and Superior throughput with minimum interference on the foreground VM services
33	SHA-1 Hash based on byte comparisons, categorizes the blocks	Server	Post	Local block metadata	Block	Variable	Fingerprints of file chunks	Time required for the deduplication of VMI and storage of VMI and metadata

The performance analysis environment is discussed in Table 3 in terms of cloud software used, type of cloud, number of images in the cloud dataset, and user level involvement for accessing this database. Table 4 gives the advantages and disadvantages of the corresponding method.

Table 3. Cloud type/ environment Parameters for non-secure Image Deduplication techniques

Paper	Cloud/OS – Cloud software	Cloud type	Number of cloud images	User level
1	VM - Amazon EC2	Private	584 VMI	Single
5	Aliyun - largest cloud of China and ISCAS – own cloud	Aliyun-Public ISCAS - private	Aliyun-Variable ISCAS- 584	Cross
6	Own cloud on a PC	Private	11 VMI	Cross
7	Openstack	Private	35 VMI	Cross
19	PeerSim P2P simulator	Private	Variable- max 30	Cross
22	Openstack	Private	102 VMI	Cross
30	OpenStack image registry with a standard configuration of 2GB memory and 10GB hard disk in CloudSim simulator	Private	4 types of virtual images - VDI, VMDK, VHD, Raw, qcow2 images in total 2,426,552,114	Cross
31	Own configuration on Ubuntu 14.04 (64-bit)	Private	10 VMI for Operating system	Single
32	Own setup with 9-servers and 16 desktops as clients running Ubuntu 12.10 64 bit with Linux kernel version 3.5.0-17	Private	276 VMI	Cross
33	Own configuration on Ubuntu 14.04 (64-bit)	Private	10 VMI for Operating	Single

Table 4. Advantages and disadvantages of non-secure Image Deduplication techniques

Paper	Advantages	Disadvantages or Limitation
1	Reduce the virtual machine image deduplication backup time	Slight storage space waste. 2 groups can have a high sample hit rate so when done in local deduplication, it can lead to duplicated blocks again
5	Work in both VHD and raw formats; accelerate the backup process	Focuses on preprocessing phase than deduplication phase; little increment of disk space usage
6	Deletion rate for image groups which have the same version of operating systems, but different versions of software applications is up about 58%	No backup system or Rapid indexing method
7	Uses the memory filter to reduce the overhead of disk index; improves the locality of data by centralizing fingerprints in disk to achieve a higher IO throughput rate with the limited memory occupancy rate	Optimization of image download process not clearly given
19	30% performance gain. Image blocks are trades in two swarms. It also deals with hash collisions	Lacks real-world environment
22	Parallel deduplication, deduplication cache and reduced memory	Cloud platform is distributed environment, but ScaleDFS is single storage partition-based deduplication
30	Very good overall reduction in image storage space and total migration time are achieved when compared with the existing image management systems	The reduction in size is dependent on the dataset and the applications running on the VM
31	Reduction in the metadata storage overhead and the number of hash computations thereby a smaller number of comparisons will be made so that overall deduplication time is reduced for the VM disk images	Dataset used in not standard
32	Parallelization of compute-intensive chunking and fingerprinting, and the I/O-intensive fingerprint indexing will speed up the deduplication process	Setting up of a cluster of deduplication servers is a costly investment
33	DedupCloud minimizes the number of hash value computations and comparisons within similar categories by using byte comparison technique	Dataset used in not standard

NON-SECURE IMAGE DEDUPLICATION FOR PIXEL IMAGE TYPE

A High-precision duplicate image deduplication approach uses the 1-norm of gray block features of images to construct B+ tree index, and then detects the possible similar images by range query [2]. It compares the number of same elements in two images edges information. The fuzzy comprehensive evaluation method is used to select duplicate images by finding the centroid image. The size ratio of deduplicated images and total images is 9.7%.

Cloud-scale image compression through content deduplication deals with combating the issues faced with storing storage costs with exponential increase in data [4]. It presents an image compression technique, which takes advantage by compressing each individual image with GIST nearest neighbor to overcome the scalability state-of-art issues.

Image deduplications check on massive image file storage that includes distributed database and file system is discussed in [8]. It uses MD5 based signature on features of binary image stream instead of file –level or block level fingerprint check.

To reduce storage space, the authors of [9] focuses on the Haar wavelet decomposition and Manhattan distance to select image duplication. When the number of same elements between two collections is greater than or equal to the preset threshold t, they considered the two images are duplicate images.

Deduplication of electricity bills is done using the content-based image retrieval with block truncation coding [10]. This is used to categorize pictures of the electricity bills and blocks of images with the same sizes are clustered together. Each cluster is checked for duplicates, and they are a part of a big block.

Deduplication image middleware detection comparison in standalone cloud database given in [11; 15] talks about techniques used in image deduplication in a standalone database. Most of the time people pay for more memory due to duplication of images. This paper shows a new framework for the early stages of image deduplication in a cloud service. 11 software taken, which are either use standalone or cloud databases. A plugin is used to detect the duplication, which is still a new topic, but mobile Cloud detection has been around from 2008. In all the software used two out of 10 is that you have high detection of duplication and those are hash and Visual similarity. The focus of the paper is to allow users to select Software and Hardware to give them a better use of the cloud services.

Large Scale Image Deduplication given in [13] deals with the problem of near Duplicate Image detection. Each duplicate in the database is linked with a Feature representation of it, what is called as a bundle. Two bundles join to form a feature of SPIHT, which is a robust technique, but become slower and gradually less accurate when the data in the database becomes larger. Maximally stable extreme regions algorithm is used for clustering as it is told to be better than the KNN means as it can also detect duplication when an image is cropped or rotated.

Authors of [17] propose a similar file extraction method where a file with high similarity is extracted. To extract similar files, average hash method is used for determining file similarity. The execution time of deduplication process can be reduced by using only similar files for comparison. Variable length blocks of files are used in this method. The average hash method is used to find the duplication of images. Morphological analysis and cosine similarity is used for the text Duplication. Results show that as the similarity percentage is increased, exact image duplicates can be determined. But the time taken for deduplication increases with increase in similarity percentage. Experimental results say that this method is very efficient to shorten the execution time.

Recognition built on vocabulary tree with indexing scheme that quantized descriptions from image key points hierarchically, which is used for image similarity indication is described in [23]. Indexing descriptor is computed for local regions. The proposed recognition method handles large number of objects for selecting one of them within the acceptable time. Local image descriptors are based on video frames extracted.

DBTP [28] i.e., Double Bytes Transport Protocol is used where double chunks are sent by the client to request for deduplication checks simultaneously, and the server responds to the deduplication requests. This scheme helps in mitigating the side channel’s risk.

DriveHQ [29] is a website developed to perform efficient image deduplication for optimized photo and video viewing. In this concept, images are divided into rectangular blocks and hash value is generated for each block using MD5.When similar images are uploaded, the images are divided into blocks and each block is checked with the stored hash values. If hash value is similar, then the images are considered as near identical images and not stored in the cloud to conserve space.

We compare the above discussed non-secure image deduplication techniques for pixel image type in the form of three table. Table 5 gives the comparison in terms of parameters related to algorithms used and perspectives of various dimensions in the image deduplication technique like Table 2. The performance analysis environment is discussed in Table 6in terms like Table 3. Table 7 gives the advantages and disadvantages of the corresponding method.

Table 5. Comparison of non-secure Image Deduplication techniques for pixel image type

Paper	**Matching Algorithm**	**Place**	**Processing time**	**Feature-based**	**Data level**	**Block size**	**Optimization objective**
2	1-norm of gray block features to construct B trees	Server	Inline	Hybrid	Block	nxn Image blocks	Duplicate images retrieval precision and deduplication accuracy
4	GIST nearest neighbor for compression	Server	Post	Local	File	NA	Image compression rates and reducing computational effort
8	MD5	Server	Post	Local	Binary stream features	Fixed	Optimization of massive image files storage
9	Manhattan distance and Haar wavelet decomposition	Server	Inline	Local	File	Fixed size file	Higher deduplication ratio, deduplication accuracy
10	Block truncation code	Client	Post	Local	Block	Variable	De-duplication process speed
11	Depend on the existing study technique	Server	Post	Local	File	NA	Storage space reduction
13	SIFT And Maximally Stable Extremal Regions (MSER),	Server	Post	Local	File	NA	Increased deduplication accuracy and performance
15	Deduplication image detector software of existing study or plugin for cloud storage	Server	Post	Local	File	NA	High-precision image deduplication
17	Average hash method, File similarity determination	Server	Post	Local	Hybrid	Fixed/ variable	Minimization of execution time for deduplication
23	Local regions indexing descriptors based on visual vocabulary tree	Server	Inline	Local	Block	Fixed	Improvement in retrieval quality
28	Double Bytes Transport Protocol with double chunks simultaneously	Client	Inline	Local	Block	Fixed	Mitigate the side channel's risk and achieve high bandwidth efficiency of deduplication
29	MD5	Server	Post	Global	Block	Fixed	Storage optimization

NA-Not Applicable

Table 6. Cloud type/ environment Parameters for non-secure Image Deduplication techniques for pixel image type

Paper	Cloud/OS – Cloud software	Cloud type	Number of cloud images	User level
2	Corel image database	Private Corel based	1000 PI	Single/cross
4	Canonical set	Private	Dynamic, millions	Cross
8	Own dataset	Private	Variable	Single
9	Corel image database and selected images from www.picsearch.com	Private	1000 PI	Single/cross
10	Own dataset	Private	Variable	Single
11	11 datasets – standard/own	Private/ public	Variable	Single/cross
13	Two dataset – Dataset of [23] for Accuracy and ILSVRC2010 for Performance	Private	Accuracy-10200; Performance- 1.2M	Cross
15	11 datasets – standard/own	Private/ public	Variable	Single/cross
17	Own cloud on a PC	Private	90 bmp images	Cross
23	Own setup	Public	40000 images of popular music CD's	Cross
28	Python 3.7.6 platform and MySQL database	Private	Variable	Cross
29	Own setup	Private	optimized photo and video viewing	Single

Table 7. Advantages and disadvantages of non-secure Image Deduplication techniques for pixel image type

Paper	Advantages	Disadvantages or Limitation
2	Reduces workload of users. The fuzzy comprehensive evaluation allows the procession of selection of centroid images by visual reference	The algorithm is unable to work on images which have been rotated, edited, blurred, or have a watermark excreta
4	Image processing rates reduces the effort used for computation by at least one order of magnitude	Ideal Canonical set is not constructed
8	Signature generation and uploading speed is improved and offers an optimization to massive image files storage	Massive image file storage distributed database without considering its deficiency
9	The proposed approach can achieve higher deduplication ratio and deduplication accuracy by setting suitable thresholds	Methods can't be used for images with similar structures
10	A single instance of the image in the database avoids confusion	Error due to entire data compression at one time
11	Evaluation of existing software is given in detail	Pilot test using standalone dataset is performed based on existing image deduplication detector
13	Method can be used even when images are cropped, rotated, or edited	Size of visual word affects performance – too small may give false results and too large will be impossible to match in one SIFT mapping
15	Deduplication image detector such as plugin, middleware or software used for deduplication	Compares standalone image deduplication detector
17	Both duplication and execution time was reduced	Most of discussion is related to text files not images. The time taken for deduplication increases with increase in similarity percentage
23	entropy weighting of the vocabulary tree is defined with video independent of the database.	Describes only retrieval process
28	DBTP implements two-side privacy to avoid side channel attack	The deduplication ratio is a little reduced compared to existing methods
29	Images are divided into blocks and hashes are generated so that duplication check can use these stored hashes and detect near identical images	Does not work for exact duplicates

SECURE IMAGE DEDUPLICATION

Secure image deduplication through image compression in the cloud storages embeds partial encryption to ensure security against a semi honest CSP and unique hashing to identify identical images into SPIHT compression algorithm [3]. Image compression followed by encryption and hashing in sequence reduces the computational overhead, resources, and metadata to be stored.

Authors of [12] discuss how cloud services have had an immense improvement in this year in terms of Secure Image Deduplication. Due to this great development in the services, many people have started storing data, which may also be redundant. Image duplication is necessary to save cost and space. This research has also used encryption called convergent encryption, which is got, by using the Hash Function on the image data the data is encrypted and decrypted with the same keys and hence the duplicates of the image will produce the same cipher text, by recognizing the duplicate of the cipher text, the image duplicates also found.

Secure Image Data Deduplication through Compressive Sensing given in [14] presents a scheme by comparing the Compression Sensing (CS) and the SPIHT technique for image deduplication according to their experimental results it is shown that the CS Technique is more efficient and has more security than the other methods. They have also further studied that this technique can be used in video duplication as well since videos also take up a lot of data space in the cloud.

The authors of [16] discuss an approach where client side takes an image, compresses it using the SPIHT compression, and partially encrypts it. It also takes the hash value of the image, and the user then uploads only the hash on to the server and the server side checks if the hash value is the same as the previous values. If it is not, then it stores the hash value or it removes the hash to eliminate redundancy.

An efficient approach towards image deduplication using Watson proposes a cost-efficient method of image duplication which has proven to reduce the storage of cloud services by one third its uses of WATSON and a MATLAB SSIM algorithm to do so [18]. In this technique when the user uploads an image it is sent to the WATSON visual where it image is given a tag and the image with the highest tag is sent to the database and is checked if other tags with the similar name is told. If it is so it is, then sent to the MATLAB SSIM to check if the images are similar or not. If the images already stored in the database, then it will not be stored again in the clients ‘profile in the Cloud Service or image is uploaded on the cloud servers and the user details are updated.

Client-Side Secure Image Deduplication of [20] uses a dice protocol, which finds image deduplication in its block level. This research concludes that with all their experimental results images, which are more like each other, have smaller number of blocks in their storage. This however does not show what happens to images, which have been cropped, are in different lighting or have scaling and any other kind of Editing done on them. It also does not deal with file, which has been compressed into different formats.

Data outsourcing model of [21] uses file level as well as block level deduplication using Dekey convergent key management scheme. A user computes and sends the block tags to the cloud server, which stores only unique tags. The stored tags are informed to the client so that it can secure it and resend back to server.

The indexed information of this secured block is maintained at client also for future access.

Secure data deduplicationusing radix trie and bloom filter discussed in [24] used existing hash-based deduplication technique. It starts with convergent encryption to avoid leakage of data followed by three stages – authorization deduplication using role re-encryption process, proof of ownership and role key update. Roles and keys are mapped with radix trie. Data updation and retrieval of ownership verification is done using bloom filter.

BDKM [25] is a blockchain based approach to ensure confidentiality of outsourced data and reliability of Convergent Key (CK) management is enhanced by adopting an oblivious pseudorandom function to generate the randomized CK. Data reliability in BDKM is achieved by dividing the CK into segments and distributed to blockchain. This work can be extended by employing blockchain to implement a secure and efficient integrity verification on the data deduplication, where a user can verify the integrity of other users' data without knowing any information about the data.

Multistage for coarse to fine deduplication is proposed in [26]. The global features are comparing to find the duplicate initially followed by local features if no match found. Fine deduplication is applied using SHA 256 based Merkle hash tree. Local and global features work at file level while hash tree works at block level. The database is maintained for each file on dataset consisting of global features, local features, and hash tree details.

Authors of [27] propose an in-line block matching-based data deduplication scheme with dynamic user management. Users encrypt their data using convergent encryption. Server uses in-line block matching protocol to generate unique proof by calculating the group key and re-encrypts the file using the group key. Another user uploading the same file will verify the proof against the server and re-encrypts using a new group key. Contents of the file remains confidential and even the server will not be aware of the file contents. Ownership list is maintained, and access control techniques are employed to prevent the access of cipher text from unauthorized users, cloud servers and adversaries. The analysis of the proposed scheme shows that the computational time, communication, and storage overhead is reduced when compared with the existing deduplication schemes.

SEDS [34] scheme is proposed to provide a secure server sided data deduplication scheme for storing data in the cloud. This scheme generates constant size ciphertext, which is independent of the number of key servers, and cloud server performs proxy re-encryption to prevent semi-honest proxy server to transform the ciphertext. This scheme supports both intra-Key Server and cross-Key Server duplication check. Experimental analysis proves that the scheme is efficient compared to previous schemes with respect to computation and communication overheads and security.

We compare the above discussed secure image deduplication techniques in the form of three tables. Like Table 2 and Table 5, Table 8 gives the comparison in terms of parameters related to algorithms used and perspectives of various dimensions in the image deduplication technique. Similarly, the performance analysis environment is discussed in Table 9 wherein additional parameters are added named as security provisioning protocol used in addition to deduplication techniques. Table 10 gives the advantages and disadvantages of the corresponding method.

DISCUSSION AND CONCLUSION

Image deduplication for duplicate check helps to reduce the communication and network transmission cost in the cloud environment. Most of the techniques work towards reduction of algorithm complexity through smaller hash size generation functions. We observed that image deduplication is either possible on pixel-based images or virtual machine images. There is no universal technique, which can be applied on both types of images. We presented taxonomy and classification of existing image deduplication related articles, which clearly shows that image.

Table 8. Comparison of Secure Image Deduplication techniques

Paper	Matching Algorithm	Place	Processing time	Feature-based	Data level	Block size	Content type	Optimization objective
3	SPIHT compression, partial encryption, and hashing	Hybrid	Inline	Global	File	NA	PI	Ensure security against a semi honest CSP and compressed image deduplication
12	attribute-based encryption	Hybrid	Post	Global	File	NA	PI	Confidentiality, Privacy protection and completeness
14	CSP - SHA basedduplicate images removal	Hybrid	Post	Global	File	NA	VMI	Ensure security against a semi honest CSP and optimize storage space
16	Robust image hashing based on SPIHT	Server	Inline	Local	File	NA	PI	Ensures data security against a curious and semi truthful CSP or any malicious user
18	WATSON and MATLAB SSIM algorithm	Server	Post	Local	File	NA	PI	Reduction in time required to perform deduplication
20	Dual Integrity Convergent Encryption protocol	Client	Inline	Local	Block	Fixed-	PI	Optimal block size determination for hashing and optimize storage space
21	DCT compression and convergent encryption	Hybrid	Inline	Local	Hybrid	Fixed - 8x8	PI	Ensure data security and optimize storage space
24	Radix Trie with Bloom Filter (SDD-RT-BF), hash function	Client	Inline	Local	Block	Fixed	PI/ Audio	Maximizing deduplication rate and ensuring security
25	Distributed Blockchain, SHA256, RSA	Client	Inline	Global	File /Block	Fixed	PI / text files	Achieve secure and reliable Convergent Key management and resistance to the brute-force attack and collusion attack launched by the external adversaries
26	Merkle-Hash and Image Features	Client	Inline	Local and global	File/Block	Fixed	PI	Ensure data security and optimize storage space
27	Guillou-Quisquater identification protocol with dynamic ownership management, Convergent encryption	Client, group Key server	Inline	—	Block	Fixed	PI / text files	Reduce network traffic and storage. Better ownership management
34	Convergent encryption and server re-encryption	Server	Inline	—	File	NA	PI / text files	Better performance of deduplication algorithm

NA- Not Applicable

Table 9. Cloud type/ environment Parameters for Secure Image Deduplication techniques

Paper	Cloud/OS - Cloud software	Cloud type	Number of cloud images	User level	Security provisioning Protocol
3	MATLAB environment dataset	Private	252 own set	Single	Partial encryption
12	No specific cloud	Any	Variable	Cross	Convergent encryption
14	MATLAB environment dataset	Private	6	Cross	semi honest CPs
16	MATLAB environment dataset	Private	10 PI	Single	Partial encryption
18	WATSON and MATLAB	Private	Variable	Cross	Password Protection
20	Own JAVA based environment	Private	30 PI	Cross	Dual Integrity Convergent Encryption
21	Not given	Private	Not given	Cross	Convergent encryption
24	Java on Amazon EC2 serve	Public	Variable	Single	SHA-256 with radix trie and bloom filter
25	No specific cloud – own index server	Private	Variable	Cross	SHA 256, RSA for encryption
26	No specific cloud – own index server	Private	Variable	Cross	SHA 256 for Merkle hash tree
27	Own server	Private	Not given	Cross	Convergent encryption
34	Own set up with multiple servers	Private	Variable	Cross	Convergent encryption and server re-encryption

Table 10. Advantages and disadvantages of secure Image Deduplication techniques

Paper	Advantages	Disadvantages or Limitation
3	Save monumental amounts of computational time and resources; Can find duplicate images even when images are extremely similar and compressed	Experimentation results are not derived in a real Cloud Service setting
12	In the paper ensure privacy protection and confidentiality	The steps done by the client are too many
14	Efficient compression scheme makes the CSP store less data	Small set of testing data with small image size
16	Great combination of analysis and security for storage	Only same images can be deduplicated. No proof of Storage protocols
18	Cloud storage space usage after deduplication has been reduced up to one third. Cost-effective	Image is only removed if user has last access to it or only the image details are hidden from the user
20	Reduce communication and bandwidth cost	Lacks real-world environment and diverse image dataset
21	DCT compression reduce storage space	small encoding/decoding overhead
24	Client-side deduplication and Tag consistency preservation with Fault tolerance	other queuing techniques and lightweight cryptographic algorithms could be used to improve performance
25	Blockchain ensures data reliability and secure key management	Secure against the collusion attack with a limited overhead and blockchain can be extended to verify integrity of other users' data without knowing the details
26	CNN is used to compare global and local features of stored images in the database with that of the incoming image and then additionally Merkle hash is used to check for duplicates	Database storage is increased for multiple level comparisons
27	File integrity is achieved by using convergent encryption, in-line block matching protocol and group key management that hides file contents from unauthorized users, cloud servers and adversaries	Generation of group keys and re encryption for subsequent uploads of the same file by other users is the overhead
34	Ensures data confidentiality, possession proof, resistant against tag inconsistency attack, cross-key server duplication check and scalability	The scheme involves multistage key generation and encryption, the process is slower. Cloud server performs proxy re-encryption which is an overhead

Deduplication has considerable potential towards efficient cloud storage usage. The proposed taxonomy has proved a convenient means of grouping the available image deduplication research and giving insight on its contribution in terms of standard features supported in the image deduplication algorithms, cloud environment, advantages, and limitations. This survey explores published research works in greater depth related to the exploitation of features of the technique used for the deduplication check. The existing image deduplication techniques neither use standard dataset as benchmark image dataset for performance evaluation nor have standard metric for similarity computation. Authors have their own way to consider performance environment. The optimization objective of the different algorithms is also listed here for researchers to get an overview of the goal of deduplication. The identified drawbacks can be scope for future research to work further for strengthen this area.

REFERENCES

1. J. Xu, W. Zhang, S. Ye, J. Wei, and T. Huang, "A lightweight virtualmachine image deduplication backup approach in cloud environment," in *2014 IEEE 38th Annual Computer Software and Applications Conference*, pp. 503–508.
2. M. Chen, S. Wang, and L. Tian, "A High-precision Duplicate Image Deduplication Approach," *JCP*, 8(11), pp.2768–2775, 2013.
3. F. Rashid, A. Miri, and I. Woungang, "Secure image deduplication through image compression," *Journal of Information Security and Applications*, 27, pp. 54–64, 2016.
4. D. Perra and J.M. Frahm, "Cloud-scale Image Compression Through Content Deduplication," in *BMVC*, 2014.
5. J. Xu, W. Zhang, Z. Zhang, T. Wang, and T. Huang, "Clustering-based acceleration for virtual machine image deduplication in the cloud environment," *Journal of Systems and Software*, 121, pp.144–156, 2016.
6. Z. Lei, Z. Li, Y. Lei, Y. Bi, L. Hu, and W. Shen, "An Improved Image File Storage Method Using Data Deduplication," in *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 638–643.
7. J. Zhang et al., "IM-Dedup: An image management system based on deduplication applied in DWSNs," *International Journal of Distributed Sensor Networks*, 9(7), p.625070, 2013.
8. S. Youjun and Z. Daxing, "Research on deduplication technology for massive image file storage," *Computer Applications and Software*, 4, p. 15, 2014.
9. M. Chen, Y. Wang, X. Zou, S. Wang, and G. Wu, "A duplicate image deduplication approach via Haar wavelet technology," in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, vol. 2, pp. 624–628).
10. A.J. Zargar, N. Singh, G. Rathee, and A.K. Singh, "Image data-deduplication using the block truncation coding technique," in *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), IEEE*, pp. 154–158.
11. N. Yusof, A. Ismail, and N.A.A. Majid, *Deduplication image middleware detection comparison in standalone cloud database*.
12. H. Gang, H. Yan, and L. Xu, "Secure image deduplication in cloud storage," in *Information and Communication Technology-EurAsia Conference*, pp. 243–251. Springer, Cham, 2015.
13. T.Y. Wen, *Large Scale Image Deduplication*. Available: http://vision.stanford.edu/teaching/cs231a_autumn1213_internal/project/final/writeup/nondistributable/Wen_Paper.pdf
14. F. Rashid and A. Miri, "Secure image data deduplication through compressive sensing," in *2016 14th Annual Conference on Privacy, Security and Trust (PST), IEEE*, pp. 569–572.

15. N. Yusof, N.A.A. Majid, and A. Ismail, "Framework deduplication image detection assisted multimedia system using multi technique," in *2016 6th International Workshop on Computer Science and Engineering, WCSE 2016*, pp. 402–406.
16. S.P. Bini and S. Abirami, "Secure image deduplication using SPIHT compression," in *2017 International Conference on Communication and Signal Processing (ICCSP), IEEE*, pp. 0276–0280.
17. T. Koike, M.Z. Nurshafiqah, and T. Kinoshita, "Data Deduplication for Similar Image Files," in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, pp. 296–301, 2018.
18. R. Aathira and V.P. Poonthottam, "An efficient approach towards image deduplication using WATSON," in *2017 International Conference on Inventive Computing and Informatics (ICICI), IEEE*, pp. 180–183.
19. C. Lee, S. Kim, and E. Kim, "A Deduplication-Enabled P2P Protocol for VM ImageDistribution," *IEICE TRANSACTIONS on Information and Systems*, 98(5), pp. 1108–1111, 2015.
20. A. Agarwala, P. Singh, and P.K. Atrey, "Client Side Secure Image Deduplication Using DICE Protocol," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE*, pp. 412–417.
21. M.S. Soofiya and S.V. Kumar, *DCT Image Compression and Secure Deduplication with Efficient Convergent Key Management.*
22. M. Ma, *Kernel-space Inline Deduplication File Systems for Virtual Machine Image Storage*; Doctoral dissertation, Chinese University of Hong Kong, 2013.
23. D. Nistr and H. Stewnius, "Scalable recognition with a vocabulary tree," in *IN CVPR*, pp. 2161–2168, 2006.
24. S.E. Ebinazer and N. Savarimuthu, "An efficient secure data deduplication method using radix trie with bloom filter (SDD-RT-BF) in cloud environment," *Peer-to-Peer Networking and Applications,* 14(4), pp. 2443–2451, 2021.
25. G. Zhang, H. Xie, Z. Yang, X. Tao, and W. Liu, "BDKM: A blockchain-based secure deduplication scheme with reliable key management," *Neural Processing Letters*, pp. 1–18, 2021.
26. D.P. Akarsha, S. Chaudhari, and R. Apama, "Coarse-to-Fine Secure Image Deduplication with Merkle-Hash and Image Features for Cloud Storage," in *2021 Asian Conference on Innovation in Technology (ASIANCON), IEEE*, pp. 1–6.
27. V. Kanagamani and M. Karuppiah, "Zero knowledge-based data deduplication using in-line Block Matching protocol for secure cloud storage," *Turkish Journal of Electrical Engineering & Computer Sciences*, 29(4), pp. 2067–2083, 2021.
28. J. Ouyang, H. Zhang, H. Hu, X. Wei, and D. Dai, "Enhanced Deduplication Protocol for Side Channel in Cloud Storages," *International Journal of Network Security*, 23(2), pp. 270–277, 2021.
29. S. Vinoth Kumar, L. Kruthika, K. Pooja, H.J. Priyanka, and N.R. Rachana, "Image Deduplication in DriveHQ Cloud," *Journal of Computational and Theoretical Nanoscience*, 17(9-10), pp. 3895–3898, 2020.
30. N.M. Tyj and G. Vadivu, "Adaptive deduplication of virtual machine images using AKKA stream to accelerate live migration process in cloud environment," *Journal of Cloud Computing*, 8(1), pp. 1–12, 2019.
31. S. Saharan, G. Somani, G. Gupta, R. Verma, M.S. Gaur, and R. Buyya, "QuickDedup: Efficient VM deduplication in cloud computing environments," *Journal of Parallel and Distributed Computing*, 139, pp. 18–31, 2020.
32. C. Lin, Q. Cao, J. Huang, J. Yao, X. Li, and C. Xie, "HPDV: A highly parallel deduplication cluster for virtual machine images," in *2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), IEEE*, pp. 472–481.
33. S.S. Patra, S. Jena, J.R. Mohanty, and M.K. Gourisaria, "DedupCloud: an optimized efficient virtual machine deduplication algorithm in cloud computing environment," *Data Deduplication Approaches: Concepts, Strategies, and Challenges*, 281, 2020.
34. S.K. Nayak and S. Tripathy, "SEDS: secure and efficient server-aided data deduplication scheme for cloud storage," *International Journal of Information Security*, 19(2), pp. 229–240, 2020.

35. D. Reinsel, J. Gantz, and J. Rydning, "Data Age 2025: The Evolution of Data to Life-Critical," *Seagate*, an IDC White Paper 2017.
36. Q. He, Z. Li, and X. Zhang, "Data deduplication techniques," in *2010 International Conference on Future Information Technology and Management Engineering (FITME)*, pp. 430–433.
37. Kirti Ashok Tayade and G.S. Malande, "Survey paper on a secure and authorized deduplication scheme using hybrid cloud approach for multimedia data," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, IEEE, pp. 2966–2969.
38. Shieh Fatemeh, Mostafa Ghobaei Arani, and Mahboubeh Shamsi, "De-duplication approaches in cloud computing environment: a survey," *International Journal of Computer Applications*, 120, no. 13, 2015.
39. W. Xia et al., "A comprehensive study of the past, present, and future of data deduplication," *Proceedings of the IEEE*, vol. 104, pp. 1681–1710, 2016.
40. "Data deduplication in the cloud explained, part one," *ComputerWorld*. Accessed on: Dec 1, 2021. [Online]. Available: https://www.computerworld.com/article/2474479/data-deduplication-in-the-cloud-explained--part-one.html
41. "Data deduplication in the cloud explained, part two: the deep dive," *ComputerWorld*. Accessed on: Dec 1, 2021. [Online]. Available: https://www.computerworld.com/article/2475106/data-deduplication-in-the-cloud-explained--part-two--the-deep-dive.html

Received 21.02.2023

INFORMATION ON THE ARTICLE

Shilpa Chaudhari, ORCID: 0000-0001-8659-4214, Ramaiah Institute of Technology, Bangalore, India, e-mail: shilpasc29@msrit.edu

Ramalingappa Aparna, ORCID: 0000-0002-8093-916X, Ramaiah Institute of Technology, Bangalore, India, e-mail: aparna@msrit.edu

ОГЛЯД ДЕДУПЛІКАЦІЇ ЗОБРАЖЕНЬ ДЛЯ ХМАРНОГО ЗБЕРІГАННЯ / Шілпа Чаудхарі, Рамалінгаппа Апарна

Анотація. Посилення комунікацій у реальному житті спонукало до створення, передавання та цифрового зберігання великих обсягів зображень і відеоданих у хмарі. Вибухове збільшення даних віртуальних/візуальних зображень на хмарному сервері потребує ефективного використання сховища, цьому посприяє технологія дедуплікації зображень. Незважаючи на те, що властивості віртуального зображення та візуального зображення розрізняються, наявна література використовує подібний підхід для перевірки дедуплікації, що спонукало розглянути обидва типи зображень для цього огляду. Дослідження має на меті надати детальний огляд найсучасніших візуальних засобів, а також методів дедуплікації віртуальних зображень у хмарному середовищі, узагальнюючи та організовуючи їх шляхом розроблення п'ятивимірної таксономії для аналізу функцій і продуктивності з кількома категоріями, що не перетинаються, у кожен вимір. До них належать: 1) місце застосування дедуплікації; 2) виділення ознак зображення; 3) час звернення; 4) стратегія розподілу даних зображення; 5) залучення рівня набору даних користувача. Наявні методи дедуплікації зображень класифікуються на дві основні категорії залежно від того, чи передбачає цей метод захист чи ні. Порівняння методів виконано за набором функціональних і продуктивних параметрів. Поточні проблеми висвітлюються з можливими майбутніми напрямами для подальших досліджень цієї теми.

Ключові слова: дедуплікація зображень, хмарні обчислення, хмарне сховище, виявлення копій зображень.

UDC 62-50
DOI: 10.20535/SRIT.2308-8893.2023.4.10

A COMPREHENSIVE SURVEY ON LOAD BALANCING TECHNIQUES FOR VIRTUAL MACHINES

SUMAN, NITIN JAIN

Abstract. Cloud computing is an emerging technique with remarkable features such as scalability, high flexibility, and reliability. Since this field is growing exponentially, more users are attracted to fast and better service. Virtual Machine (VM) allocation plays a crucial role in cloud computing optimization; hence, resource distribution is not impacted by machine failure and is migrated with no downtime. Therefore, effective management of virtual machines is necessary for increasing profit, energy-saving, etc. However, it could utilize the virtual machine resources more efficiently because of the increased load, so load balancing is more concentrated. The predominant purpose of load balancing is to balance the available load equally among the nodes to avoid overloading or underloading problems. The present study conducted an extensive survey on virtual machine placement to describe the application of prediction algorithms and to provide more efficient, reliable, high response, and low overhead VM placement. Furthermore, the survey attempted to overview the challenges in load balancing in VM placement and various ideas of state-of-the-art techniques to resolve the issues..

Keywords: virtual machine allocation, load balancing, cloud computing, overloading, physical machine, data center.

INTRODUCTION

Cloud computing is now becoming vital for hosting several IT services that provide various on-demand VR (Virtual Resources). The cloud service providers used large-scale DC (data center) with more physical machines. Virtualization is more beneficial in data centers for providing the VM comprising a software layer known as a VVM Monitor. This VMM enables the controlling of shared physical machine resources, thereby increasing VM security but accommodating multiple VM in a single physical machine remains a challenging problem. Due to this problem, there is a chance of overutilizing PM, degrading it, or wasting high-cost resources.

Further, the power consumption of cloud DC mainly occurs by physical machines, so proper VM placement associated with dynamic management greatly mitigates DC power consumption, improving the profit and throughput and preventing SLA violations. However, VM placement employs a widespread and expensive VM migration process. If improper placement occurs, it will lead to the destruction of data centre performance. Furthermore, balancing the request's workload and allocating appropriate tasks to the appropriate VM is also considered challenging. Hence load balancing is a crucial factor to be considered with the increasing requests and impulsive arrival patterns. Load balancing is the even classification of the task processed in-between more CPUs, storage devices, and network links which deliver fast service with more efficiency. This is obtained using hardware/software devices and multiple servers that appear as a computer clustering. In addition, load balancing improves the efficiency of distributed or parallel systems through load redistribution. Load balancing algorithms are classified into static and dynamic algorithms. The static algorithm is simple and needs minimized runtime overhead, whereas a dynamic system is utilized in most of the modern load-balancing approaches due to its flexibility and robustness. Likewise, there exist four kinds of load-balancing policies, which are location policies, transfer policies, selection policies, and information policies. Other such objectives of load balancing include reducing carbon emission and energy consumption, resource provisioning, avoiding bottlenecks, and achieving QoS requirements.

To overcome the prevailing limitations and obtain end-user satisfaction, high-quality and effective methodologies must be adopted to support the optimization of VM load balancing. Therefore, the study's main contribution is to provide a comprehensive survey of the existing methods dealing with virtual machine load balancing by the factors affecting the cloud computing process.

Objective:

- To analyze several virtual machine placement and load balancing techniques in the existing literature.
- To overview the prevailing challenges in load balancing in virtual machine processing and to provide a comprehensive outlook to rectify the issues.
- To outlook the recent trends for the optimization of load balancing in virtual machine allocation.

The paper is organized as follows: Section 2 deals with the predictive virtual machine placement methods with various algorithms, and Section 3 reviews prevailing load balancing techniques such as static and dynamic methods. Section 4 summarizes the advantages of load balancing in virtual machine allocation, and Section 5 overviews the performance metrics for evaluating load balancing in virtual machines. Section 6 provides the recent trends in this concept, and Section 7 deliberates on the challenges and research gaps of load balancing for virtual machines. Followed by Section 8 concludes the work.

PREDICTIVE VIRTUAL MACHINE PLACEMENT METHODS

Various predictive virtual machine placement methods are designed and suggested for the CC environment. Besides, implementing all these methods for enhancing the placement process of virtual machines by utilizing historical data. The predictive methods for the VM placement are classified as the following [1].

- Ensemble-based scheme.
- Hybrid scheme.
- Exponential smoothing predictor-based scheme.
- Dynamic programming-based scheme.
- Grey model-based scheme.
- Fractal based schemes.
- Bayesian-based scheme.
- Neural network-based scheme.
- SVM based scheme.
- Queuing based scheme.
- Markov based scheme.

- Hidden Markov Model based scheme.
- ARIMA model-based scheme.
- Regression-based scheme.

The following are a few existing virtual placement schemes that utilize the above-said algorithms. This study [2] attempted to predict resource requirements for virtual machines to improve the process of virtual machine placement. Besides, the study indicated that the time-reliable hidden Markov model replicated the properties of CPU utilization data and determined the CPU's future usage. Nevertheless, the study applied only univariate normal distribution to determine resources, whereas the multivariate normal distribution to determine multiple resources could be useful. Likewise, [3] provided a Markov predicting framework for forecasting the future under-utilized/over-utilized physical machines and preventing unnecessary and immediate migration of virtual machines. Besides, this study utilized the cloud sim toolkit evaluated using random forest and Planet Lab datasets. Finally, the current usage of the CPU of every physical machine has been compared with upper and lower thresholds for recognizing the status.

Moreover, determine the future state of physical machines by implementing the Markov model. This study [4] introduced a framework for identifying the relationship of resources amid virtual machines by utilizing ARIMA-based determinations. Further, the study analyzed resource utilization after the placements of two virtual machines on the same physical machine, and also the study named this system an affinity model. Similarly, this study [5] implemented automata for enhancing the usage of resources as well as mitigating the usage of power. Furthermore, this study considered the variations in user demands for estimating overloaded physical machines. Due to the prevention of physical machine overload, this system improves the utilization of resources, mitigates the amount of migration, and shuts the idle physical machines to mitigate the utilization of power. Finally, the study stated that this method was executed in the cloudsim toolkit by utilizing Planet Lab dataset. Nevertheless, this cannot detect underutilized physical machines.

REVIEW OF PREVAILING LOAD BALANCING METHODS

This study surveyed the literature on prevailing load-balancing methods and comprehensively reviewed certain studies. Besides, the load balancing methods are segregated into dynamic and static, based on the system's state. Load balancing is needed to improve resource utilization, reducing the completion and response time for the tasks on the cloud. This study [6] suggested a method in which it considered QoS, number of migrations as well as response time as the parameters of load balancing. Further, tasks with less priority have been transferred from one virtual machine to another when overloaded with virtual machines. This method can be improvised with other algorithms like ACO and PSO.

Static load balancing methods. The static load balancing method doesn't require knowledge of a system's current state; it requires knowledge of the system resources like processing power, storage capacity, memory, and execution time in advance. Besides, the static load balancing methods don't allow resource allocation at execution time. Also, these methods are easy to execute and implement, but they are beneficial to small networks or systems with a minimum amount of resources. On the other hand, as they don't consider the present state of the system, these methods aren't beneficial for computing systems that perform distributed computing. Moreover, they need to permit the detection of connected server machines at the execution time, thereby leading to uneven resource distribution.

Dynamic load balancing methods. Since the static load balancing techniques are not suitable for the distributed computing system, the dynamic load balancing methods are suitable in a cloud computing environment. The following are different load balancing methods, which rely on the criteria of the load balancer:

- Cluster-based load balancing.
- Task-based load balancing.
- Agent-based load balancing.
- Hybrid load balancing.
- Natural phenomena based on load balancing.
- General load balancing.
- Cluster-based load balancing.

This study [7] addressed a heuristic method for load balancing based on (LB-BC) Bayes and Clustering for overcoming the difficulties of prevailing load balancing techniques. This technique is based on Bayes' theory and has accomplished long-term load balancing. This computes the posterior probability of the physical hosts and integrates with clustering for picking an optimal host. Further, this considered the parameters like load balancing effect, standard deviation, and the number of requested tasks. Then, it was compared with the dynamic load balancing, leading to increased time and minimal standard deviation. This method only works in localized areas, but further enhancement can be made for working in a real-time environment and a wide area network. This study [8] presented a cluster-based method for improvising intercloud communication in real-time and dynamic multi-media for load balancing. This method has a two-step process. The first step is to develop the cluster to monitor the activities, handle platform difficulties, and meet the satisfactory quality of service and demands for hosts based on a hello-packet broadcast for all the servers. In the second step, it decides on transfer job requests. When this method was compared with HFA, WCAP, and ant colonies, the suggested method produced an improved response time. In addition, this method could be improvised for a real-time environment, in which the intermediate nodes are congested, and owing to reduce the data loss because of congestion by utilizing communication jobs instead of computation jobs.

This study [9] presented a cluster-based load-balancing method for overcoming load distribution issues. Besides, this integrated the concept of KUHN and genetic algorithm and created a task allocation strategy by grouping the tasks into clusters and distributing them in a cooperating node. As a result, this method provided improvised task distribution and response time among data center nodes. Similarly, this study [10] created a hierarchical model to self-schedule the schemes for improving the scalability and load balancing of the cloud system. Besides, this method can extract in a heterogeneous and homogeneous environment. In addition, this study has implemented the schemes on a large scale by utilizing various computation applications. Finally, the outcomes of the study depicted improvised scalability and overall performance, as well as decreased communication overhead. The further analysis deals with a testing algorithm for large-scale loops and clusters with dependencies.

Task-based load balancing. This study [11] presented a network-aware task placement method for reducing task completion time, data cost, and transmission time. The study stated that the three challenges faced by tasks are the availability of resources dynamically changes resulting in access over time; data fetching time relies on the task's location and size; the load on the path significantly impacts the data access latency. Therefore, the study must consider loading over the path during scheduling to minimize this latency. The study's outcomes depicted that the suggested method has significantly reduced the task's completion time and increased resource utilization.

This study [12] suggested a scheduling technique for reducing the resource competition between high device load and tasks based on the weighted random scheduling method. The tasks are assigned by considering parameters such as communication delay, time, and cost. Besides, the study was analyzed with MATLAB software by utilizing workflow for generating the dataset. Also, the study analyzed the dataset, which included a large set of tasks with transmission delay, cost, and time. Moreover, the study considered device dependency, task arrival time, and task structure. The study's outcomes depicted that multiple schedules have seen improvement in parameters like execution cost and task completion for the devices. Nevertheless, it still needs to calculate the optimal value for parameters that could be improvised in further analysis.

Agent-based load balancing. The multi-agent-based load-balancing framework helps increase resource utilization [13]. This executed both the receiver originate method, as well as the sender, originated method for reducing the waiting time of tasks and also for assuring SLA. This method incorporated the agents like NA (Negotiator Ant) agent, DCM (Datacenter Monitor) Agent, as well as VMM (Virtual Machine Monitor) Agent. Among these, the virtual machine monitor agent supports every virtual machine in the system and retains the information on bandwidth, CPU, and memory by utilizing virtual machines for monitoring the load. Besides, the datacentre monitor agent executes information policy by utilizing the available information from the virtual machine monitoring agent and categorizing the virtual machines relying upon various characteristics. Also, this initiates the negotiator and agent that moves to various other data centers for identifying the available virtual machines' status. From the experimental analysis, the study stated that the suggested method was more effective, improving the response time and reducing the makespan time.

The (SVLL) selection of virtual machines with the least load balancing technique for the distribution of tasks increased the cloud computing performance [14]. This model computes a load of every virtual machine and assigns tasks to evaluate based on the virtual machine's load rather than the number of tasks assigned to virtual machines. Besides, the study implemented the SVLL method with various task scheduling methods like shortest job first and first came first serve methods, in which the outcomes of the study denoted that the suggested method has improvised in total finishing time and total waiting time. In addition, this method was employed with basic task scheduling methods for better results.

This study [15] developed a load-balancing method by integrating round-robin features and shortest-job-first scheduling algorithms. This method stores long and short tasks in separate queues and utilizes dynamic task scheduling quantum to balance waiting time among the tasks. Besides, this study has taken into account the issues of starvation as well as throughput. Also, they executed the experiment on the cloud tool. As a result, the experimental analysis showed that response time, waiting time, and the turnaround time was reduced. In addition to that, long-task starvation was also minimized. Nevertheless, the task quantum was not efficient in balancing the tasks, but it could be improvised in further analysis.

The hybrid load-balancing method. This study [16] employed a hybrid algorithm for optimizing the system's performance by integrating throttled and round-robin load balancing methods with a service-proximity broker and performance-optimized service broker algorithm. Besides, the study suggested one load balancing and three service broker methods. The study denoted them as CA (Cost Aware) and LA (Load Aware) algorithms for high utilization of resources. However, although the LA algorithm offers low processing time, it can generate high costs, whereas CA reduces cost. Moreover, the service broker algorithm decides on the server to users' requirements, which might increase cost or processing time.

In contrast, the service proximity algorithm decides on the data center near to client's region. Finally, the study integrated all the algorithms, and the outcomes of the study denoted that response time and processing time have significantly reduced. Nevertheless, further analysis deals with the improvisation of system performance. The development of an efficient CLB (Cloud Load Balancing) framework is needed to overcome the server failure response in the event of several user requests. Several studies have developed a framework that considers the loading and server processing for minimizing the server problems for handling various computation requests. Also, they presented a load-balancing method for virtual and physical web servers to preserve the information regarding computing power, priority, and server loading. Even though this framework provides high scalable performance, it can increase response time.

COMPARATIVE ANALYSIS OF STATIC AND DYNAMIC LOAD BALANCING TECHNIQUES

Table provides a comprehensive comparative analysis of the existing load-balancing algorithms.

Comparative Analysis between The Existing Load Balancing Algorithms

S. No	Type of Load Balancing algorithm in VM	Load Balancing Algorithm	Parameters enhanced	Merits	Demerits
1	Static	Weighted-round robin algorithm [17]	Waiting and response time	Utilize all resources in a balanced manner. Ensuring fairness in every allocation	Execution time Prediction is not possible. High Migration time
2		Opportunistic load-balancing algorithm [18]	User discomfort cost and reduction	The end-user achieves better accuracy and comfort maximization	Comfort maximization might lead to raised costs and energy
3		Software-Defined Networking based load-balancing algorithm [19]	Cost, response time, and scalability	Effective user request processing	Increased energy consumption
4	Dynamic	Ant colony algorithm [20]	Makespan, response time, scalability	Good scalability, Fault tolerance, and obtaining load balancing for Complex networks.	High power consumption. Less throughput
5		Deadline-constrained based dynamic load-balancing algorithm [21]	Task rejection ratio, makespan	Increases the utilization ratio	Increased consumption of cost
6		Honey-bee foraging algorithm [22]	Response time, throughput	Less waiting time and Increased system diversity	High response time Less throughput

BENEFITS OF LOAD BALANCING IN VIRTUAL MACHINES

Ideally, these solutions can be implemented when performing the placement of virtual machines. Decreasing the number of physical machines as well as consolidating virtual machines could be utilized for solving cloud-spot issues. Reducing the migrations of virtual machines by predicting future workloads will prevent unnecessary migrations of virtual machines. Future pages could be identified by mitigating transmitted pages by properly predicting the workload of applications. Consequently, the number of transmitted pages could be diminished in the pre-copy approach.

The load balancer offers flexibility for balancing the server's workload by traffic distribution across multiple servers. Further, load-balancing targets mimic a software infrastructure via Virtualization. This runs physical load-balancing software on VM. In addition, availability, performance, scalability, and reliability are the major metrics of load balancing.

Availability. The mechanism of load balancing assures an efficient offer of service. Moreover, the loads will be effectively distributed in terms of server unavailability.

Performance. An effective load balancing provides cloud applications as well as cloud services for responding faster when compared to the average completion time. In addition, execution time is also decreased via effective compression methods and catching mechanisms.

Scalability. The major benefit of the load-balancing technique is that some servers can be easily included without any disturbance, and the applications can be smoothly performed via the load-balancing servers.

Reliability. The reliability of cloud services was secured by the redundancy of servers in which the applications could be hosted. Even in failure cases, the cloud-serving resources will function, and its services will be redirected to other locations in the cloud.

PERFORMANCE METRICS IN THE EVALUATION OF LOAD BALANCING FOR VIRTUAL MACHINES

Various virtual machine load balancing metrics are present for assessing load balance performance. These metrics were reflected in diverse task scheduling behavior. The following are the load balancing metrics.

Load variance. Consider that there exists *n* number of hosts in the data center. The usage of host *i* can be expressed as $U(host_i)$, whereas the average usage of every host can be calculated as

$$avg\,(U_t) = \frac{1}{2}\sum_{i=1}^{n} host_i .$$

Makespan time. Makespan time is known as the longest-processing time on every host. Also, it is a normal criterion for accessing scheduling algorithms. Retaining load balance is for shortening the makespan time.

Overloaded hosts. The overload threshold can be denoted as $T(U_t)$, for *n* number of hosts, the host utilization can be expressed as $U(host_i)$, and the overload hosts is expressed as the following, $Num(T(U_t)) \le U(host_i)$.

Throughput. Throughput deals with system performance. A maximum number of tasks are executed to accomplish high performance within the minimal completion time.

SLA violations. Similarly, this also deals with the performance of the system. The virtual machines can't fetch adequate resources from the host, so the host isn't well-balanced. Thus, SLA violations must be reduced.

Turnaround time. Turnaround time is defined as the time systems take from the request submission to a response from the server. And turnaround time can be calculated as

$$\text{Turnaround time } = C_t - C_T .$$

From the above equation C_t refers to completion time, and *GT* refers to generation time.

Overhead. Generally, overhead occurs because it increases the communication cost or takes more time to migrate from one virtual machine to another. Good load-balancing algorithms will decrease the overhead.

Resource usage. Good performance usually deals with the proper resource usage among nodes. This will be beneficial for measuring if the nodes are underloaded or overloaded.

Fault tolerance. This enhances the systems such that the single failure point doesn't impact the entire system. Besides, the load balancing algorithm must be designed in a way where the failure of one node must not affect the system.

Response time. Generally, response time is the time taken by load balancing techniques to users. Lesser response time indicates better system performance. Therefore, load balancing will be more beneficial for the entire cloud by decreasing the response time of cloud servers and task scheduling issues; the following articles discuss the response time in virtual machines.

This study [23] suggested TMA (Throttled Modified Algorithm) improves the response time of virtual machines on CC (Cloud Computing) to improvise a performance. Besides, this study simulated the suggested method with the clouds tool; the evaluated outcomes showed improved processing time and response time.

In this study [24], a firefly load balancing technique was utilized to solve the load imbalance problems in a cloud server to enhance the learners' user experience. The suggested method needs a cloud-server mapping method for various virtual machine methods, ensuring the users receive the content without delay. From the experimental analysis, the study stated that, compared to the existing method, the suggested method showed less response time.

RECENT TRENDS OF LOAD BALANCING IN VIRTUAL MACHINE ALLOCATION

This study [25] suggested that response time was similar to execution time in every task, and this parameter should be minimized. This determines the virtual machine status based on the current load. Later, the tasks are eliminated from the machine with additive load, which depends on the virtual machine's condition. Finally, it will be transferred to the appropriate VM, which is the criteria to assign tasks to virtual machines based on the least distance. The outcomes of the cloudsim tool evaluation showed that response time was improved compared to existing algorithms. Additionally, the degree of load imbalance has also seen some improvements.

The main aim of task scheduling incorporates scheduling resources and reducing the schedule's objective. This study [26] suggested a mean grey-wolf optimization technique to enhance the system's performance and reduce scheduling problems. The primary objective of this study is to reduce energy consumption and makespan time. This was evaluated by utilizing the cloudsim tool. The study showed that the suggested algorithm had better results than the prevailing methods.

This suggested method in this study [27] attempted to avoid SLA violations via power optimization and optimal cloudlet by reducing the migrations of virtual machines. Besides, the SLA reduction system incorporated three parts a scheduling algorithm, a MinVM scheduling algorithm, and a credit-based virtual machine migration algorithm. When considering the scheduling algorithm, it efficiently schedules the cloudlets to VMs based on the host's processing time. Likewise, the MinVM scheduling algorithm schedules the cloudlets to VMs based on counts of cloudlet allocation to every virtual machine. And the credit-based algorithm utilizes the virtual machine's credit to take virtual machine migration.

CHALLENGES AND RESEARCH GAP

The most challenging task in virtual technology is virtual machine placement on the physical machine under optimal conditions in cloud-data centers. Further, the virtual machine placement can result in managing resources and preventing the wastage of resources. Minimizing energy consumption, cost reduction, utilization of resources, and presentation of best QoS are significant challenges in the cloud computing environment. Since only a few studies focus on privacy and security issues, in further analysis, security is a crucial factor that must be focussed on. Besides, attackers can steal the secrets from other tenants by utilizing side-channel attacks based on shared resources since the virtual machines from various tenants might be located at one physical machine, thereby threatening data security in a cloud computing environment. The following are certain limitations that should be considered,

- The forecasting approaches employed in predictive virtual machine placement schemes could be enhanced to better deal with non-linear and linear loads.
- Moreover, the predictive virtual machine placements in the multi-cloud and multi-site cloud environments must be studied for further analysis.
- Even though dynamic power management can be implemented to improvising DCs energy efficiency, only a few studies have suggested this approach in their literature.
- One of the significant problems avoided by various studies is DDoS attacks that could be originated from malicious virtual machines by uplifting the resource demands and introducing several unnecessary virtual machine migrations.
- The integration of predictive virtual machine placement methods with prevention and intrusion detection systems must be investigated to recognize the true demands and increase the DC's security.

- Studies must design and apply low-overhead placement methods in developing technologies like mobile and cloudlets in the future.
- In future studies, context-aware virtual machine placement must be designed for the environment, like predicting mobile patterns, vehicular CC, and connectivity problems.

In recent years, cloud computing has seen rapid growth and advanced research in computation and data based on practical and theoretical aspects. Nevertheless, cloud computing researchers face several problems in which load balancing is more challenging and needs special attention. Besides, issues like user QoS (Quality of Service) satisfaction, virtual machine security, resource usage, and virtual machine migration must be considered to find a feasible solution to improve resource utilization. Additionally, various problems like the migration of virtual machines, resource utilization, QoS satisfaction, and migration of virtual machines need equal attention for finding the optimal solution to enhance the optimal solution to improve the utilization of resources.

The following are certain load-balancing problems.

Geographically distributed nodes. Generally, the data centers in the cloud are geographically-distributed. In these data centers, for effective system execution according to the request of users, the spatially distributed nodes were treated as a single location system. Besides, certain load balancing methods were designed for a small area. For example, they don't consider communication delay, network delay, and the distance between distributed resources, users, and computing nodes. Nevertheless, the nodes situated at various locations are challenging since these algorithms are unsuitable for these environments. Therefore, load-balancing techniques for distantly located nodes must be considered [28].

Migration of Virtual Machines. Virtualization allows for the creation of numerous virtual machines on one physical machine. As a result, virtual machines are generally independent and possess various configurations. Besides, if the physical machine is overloaded, certain virtual machines must migrate to a distant location using the virtual machine migration load balancing method [29].

Heterogeneous Nodes. During earlier research in load balancing, several studies have theorized about homogeneous nodes. But, usually, in cloud computing, users' requirements dynamically change, which needs executing time for efficient resource utilization and decreasing response time. Thus, introducing an effective load-balancing method for the heterogeneous environment is more challenging [30].

Storage management. Cloud storage solved the issues of the conventional storage system, which required higher hardware costs and personnel management. Further, the cloud allows users to store data heterogeneously without access issues as there is a rapid increase in cloud storage, data replication for data consistency, and effective access. However, because of duplicate storage policies, full data replication is ineffective. The partial replication could be adequate. However, there are certain issues in the dataset's availability, and there might be increased complexities in load balancing methods [31].

Scalability of the load balancer. The scalability and on-demand availability of cloud services allow users to access the services for rapidly scaling up or scaling down. Therefore, a load balancer must consider rapid variations by system topology, storage, and computing power to efficiently facilitate these variations [32].

Complexity of algorithms. In a cloud computing environment, usually, the algorithms must be easier and simple to implement. Besides, complex algorithms may diminish the efficiency and performance of cloud systems [33].

CONCLUSION

In general, a cloud indicates a distinct IT environment designed for the proper functioning of remotely providing scalable and measurable IT resources. The paper's main objective is to consolidate the prevailing VM placement and load-balancing methodologies. Further various challenges to the enhancement of effective VM and load-balancing algorithms are also discussed. This survey lets the users look at recent trends in VM placement and load balancing, enabling them to frame an effective research methodology with maximum profit and minimum cost.

Declaration. I confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere.

Acknowledgments

- None.

Funding

- Any organization/institute/agency did not fund this research work.

Competing Interests

- None of the authors have any competing interests in the manuscript.

Availability of data and material

- Not Available.

Code availability

- Not Available.

Compliance with ethical standards

Ethical statement

- No human participants or animals are involved in this research.

Consent statement. I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report) have been allowed to review the final manuscript and have provided written consent to publish.

REFERENCE

1. M. Masdari and M. Zangakani, "Green cloud computing using proactive virtual machine placement: challenges and issues," *Journal of Grid Computing*, pp. 1–33, 2019.
2. H.L. Hammer, A. Yazidi, and K. Begnum, "An inhomogeneous hidden Markov model for efficient virtual machine placement in cloud computing environments," *Journal of Forecasting*, vol. 36, pp. 407–420, 2017.
3. S.B. Melhem, A. Agarwal, N. Goel, and M. Zaman, "Markov prediction model for host load detection and VM placement in live migration," *IEEE Access*, vol. 6, pp. 7190–7205, 2017.
4. X. Fu and C. Zhou, "Predicted affinity based virtual machine placement in cloud computing environments," *IEEE Transactions on Cloud Computing*, vol. 8, pp. 246–255, 2017.
5. M. Ranjbari and J.A. Torkestani, "A learning automata-based algorithm for energy and SLA efficient consolidation of virtual machines in cloud data centers," *Journal of Parallel and Distributed Computing*, vol. 113, pp. 55–62, 2018.
6. K.R. Babu and P. Samuel, "Enhanced bee colony algorithm for efficient load balancing and scheduling in cloud," in *Innovations in bio-inspired computing and applications*. Springer, 2016, pp. 67–78.
7. J. Zhao, K. Yang, X. Wei, Y. Ding, L. Hu, and G. Xu, "A heuristic clustering-based task deployment approach for load balancing using Bayes theorem in cloud environment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, pp. 305–316, 2015.

8. B. Kang and H. Choo, “A cluster-based decentralized job dispatching for the large-scale cloud,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, pp. 1–8, 2016.
9. F. Zegrari, A. Idrissi, and H. Rehioui, “Resource allocation with efficient load balancing in cloud environment,” in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, 2016, pp. 1–7.
10. Y. Han and A. T. Chronopoulos, “Scalable loop self-scheduling schemes for large-scale clusters and cloud systems,” *International Journal of Parallel Programming*, vol. 45, pp. 595–611, 2017.
11. H. Shen, A. Sarker, L. Yu, and F. Deng, “Probabilistic network-aware task placement for mapreduce scheduling,” in *2016 IEEE International Conference on Cluster Computing (CLUSTER)*, pp. 241–250.
12. Y. Xin, Z.-Q. Xie, and J. Yang, “A load balance oriented cost efficient scheduling method for parallel tasks,” *Journal of Network and Computer Applications*, vol. 81, pp. 37–46, 2017.
13. S. Keshvadi and B. Faghih, “A multi-agent based load balancing system in IaaS cloud environment,” *International Robotics & Automation Journal*, vol. 1, pp. 1–6, 2016.
14. T. Aladwani, “Impact of selecting virtual machine with least load on tasks scheduling algorithms in cloud computing,” in *Proceedings of the 2nd International Conference on Big Data, Cloud and Applications*, 2017, pp. 1-7.
15. S. Elmougy, S. Sarhan, and M. Joundy, “A novel hybrid of Shortest job first and round Robin with dynamic variable quantum time task scheduling technique,” *Journal of Cloud computing*, vol. 6, pp. 1–12, 2017.
16. R.K. Naha and M. Othman, “Cost-aware service brokering and performance sentient load balancing algorithms in the cloud,” *Journal of Network and Computer Applications*, vol. 75, pp. 47–57, 2016.
17. K. Manojkumar and P. Kanagaraju, *Enhanced load balancing algorithm to reduce response time and waiting time by incorporating weighted round robin and honey bee behaviour algorithm in cloud computing.*
18. M.B. Rasheed, N. Javaid, M.S.A. Malik, M. Asif, M.K. Hanif, and M.H. Chaudary, “Intelligent multi-agent based multilayered control system for opportunistic load scheduling in smart buildings,” *IEEE Access*, vol. 7, pp. 23990–24006, 2019.
19. H. Zhong, Y. Fang, and J. Cui, “Reprint of “LBBSRT: An efficient SDN load balancing scheme based on server response time,” *Future Generation Computer Systems*, vol. 80, pp. 409–416, 2018.
20. S. Dam, G. Mandal, K. Dasgupta, and P. Dutta, “An ant-colony-based meta-heuristic approach for load balancing in cloud computing,” in *Applied Computational Intelligence and Soft Computing in Engineering*, IGI Global, 2018, pp. 204–232.
21. M. Kumar and S. Sharma, “Deadline constrained based dynamic load balancing algorithm with elasticity in cloud environment,” *Computers & Electrical Engineering*, vol. 69, pp. 395–411, 2018.
22. V. Bhavya, K. Rejina, and A. Mahesh, “An Intensification of Honey Bee Foraging Load Balancing Algorithm in Cloud Computing,” *International Journal of Pure and Applied Mathematics*, vol. 114, pp. 127–136, 2017.
23. N.X. Phi, C.T. Tin, L.N.K. Thu, and T.C. Hung, “Proposed load balancing algorithm to reduce response time and processing time on cloud computing,” *Int. J. Comput. Networks Commun.*, vol. 10, pp. 87–98, 2018.
24. K. Sekaran, M.S. Khan, R. Patan, A.H. Gandomi, P.V. Krishna, and S. Kallam, “Improving the response time of m-learning and cloud computing environments using a dominant firefly approach,” *IEEE Access*, vol. 7, pp. 30203, 2019.
25. L. Xingjun, S. Zhiwei, C. Hongpong, and B.O. Mohammed, “ A new fuzzy-based method for load balancing in the cloud based Internet of thing usingagrey wolf optiization algorithm,’’ *International Journal of Communication Systems*, vol. 33, p. e4370, 2020.

26. G. Natesan and A. Chokkalingam, "Task scheduling in heterogeneous cloud environment using mean grey wolf optimization algorithm," *ICT Express*, vol. 5, pp. 110–114, 2019.
27. M.K. Halili and B. Cico, "SLA management for comprehensive virtual machine migration considering scheduling and load balancing algorithm in cloud data centers," *International Journal on Information Technologies & Security*, vol. 12, 2020.
28. A.K. Kiani and N. Ansari, "On the fundamental energy trade-offs of geographical load balancing," *IEEE Communications Magazine*, vol. 55, pp. 170–175, 2017.
29. N.H. Shahapure and P. Jayarekha, "Virtual machine migration based load balancing for resource management and scalability in cloud environment," *International Journal of Information Technology*, pp. 1–12, 2018.
30. X. Shao, M. Jibiki, Y. Teranishi, and N. Nishinaga, "An efficient load-balancing mechanism for heterogeneous range-queriable cloud storage," *Future Generation Computer Systems*, vol. 78, pp. 920–930, 2018.
31. S. Subalakshmi and N. Malarvizhi, "Enhanced hybrid approach for load balancing algorithms in cloud computing," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 2, pp. 136–142, 2017.
32. T.G. Rodrigues, K. Suto, H. Nishiyama, N. Kato, and K. Temma, "Cloudlets activation scheme for scalable mobile edge computing with transmission power control and virtual machine migration," *IEEE Transactions on Computers*, vol. 67, pp. 1287–1300, 2018.
33. M. Xu, W. Tian, and R. Buyya, "A survey on load balancing algorithms for virtual machines placement in cloud computing," *Concurrency and Computation: Practice and Experience*, vol. 29, p. e4123, 2017.

Received 21.11.2022

INFORMATION ON THE ARTICLE

Suman Sansanwal, ORCID: 0000-0001-8485-0931, Chandigarh University, Punjab, India, e-mail: sumanphd27@gmail.com

Nitin Jain, Chandigarh University, Punjab, India, e-mail: nitin@gmail.com

КОМПЛЕКСНИЙ ОГЛЯД ТЕХНІК БАЛАНСУВАННЯ НАВАНТАЖЕННЯ ДЛЯ ВІРТУАЛЬНИХ МАШИН / Суман Сансанвал, Нітін Джайн

Анотація. Хмарні обчислення — це нова техніка з чудовими характеристиками, такими як масштабованість, висока гнучкість і надійність. Оскільки ця сфера експоненціально зростає, швидке та якісне обслуговування приваблює більше користувачів. Розподіл віртуальної машини (VM) відіграє вирішальну роль в оптимізації хмарних обчислень; на розподіл ресурсів не впливає збій машини та перенесення відбувається без простоїв. Ефективне керування віртуальними машинами необхідне для збільшення прибутку, енергозбереження тощо. Однак воно може більш ефективно використовувати ресурси віртуальної машини через збільшення навантаження, тому балансування навантаження є більш концентрованим. Переважна мета балансування навантаження — рівномірно збалансувати доступне навантаження між вузлами, щоб уникнути проблем із перевантаженням або недовантаженням. У дослідженні виконано розширений огляд щодо розміщення віртуальних машин, щоб описати застосування алгоритмів прогнозування та забезпечити більш ефективне, надійне розміщення віртуальної машини з високою відповіддю та низькими накладними витратами. Крім того, у ході роботи зроблено спробу оглянути проблеми балансування навантаження у розміщення віртуальної машини, а також різні ідеї щодо сучасних методів вирішення цих проблем.

Ключові слова: розподіл віртуальної машини, балансування навантаження, хмарні обчислення, перевантаження, фізична машина, центр оброблення даних.

ВІДОМОСТІ ПРО АВТОРІВ

Баздирев Антон Андрійович,

аспірант ІПСА КПІ ім. Ігоря Сікорського, Україна, Київ

Базілевич Ксенія Олексіївна

доцент, кандидат технічних наук, доцент кафедри математичного моделювання та штучного інтелекту Національного аерокосмічного університету імені М.Є. Жуковського «Харківський авіаційний інститут», Україна, Харків

Виклюк Ярослав Ігорович,

професор, доктор технічних наук, професор кафедри систем штучного інтелекту Національного університету «Львівська політехніка», Україна, Львів

Донець Володимир Віталійович,

аспірант кафедри теоретичної та прикладної системотехніки Харківського національного університету імені В.Н. Каразіна, Україна, Харків

Дорогий Ярослав Юрійович,

доцент, доктор технічних наук, доцент кафедри інформаційних систем та технологій КПІ ім. Ігоря Сікорського, Україна, Київ

Кіріленко Олена Юріївна,

здобувач кафедри математичного моделювання та штучного інтелекту Національного аерокосмічного університету імені М. Є. Жуковського «Харківський авіаційний інститут», Україна, Харків

Колісніченко Вадим Юрійович,

аспірант кафедри інформатики та програмної інженерії КПІ ім. Ігоря Сікорського, Україна, Київ

Кривцов Сергій Олегович,

аспірант кафедри математичного моделювання та штучного інтелекту Національного аерокосмічного університету імені М. Є. Жуковського «Харківський авіаційний інститут», Україна, Харків

Кузнєцова Вікторія Олександрівна,

кандидат фізико-математичних наук, старший викладач кафедри вищої математики та інформатики Харківського національного університету імені В.Н. Каразіна, Україна, Харків

Мартьянов Дмитро Ігорович,

аспірант кафедри систем штучного інтелекту Національного університету «Львівська політехніка», Україна, Львів

Меняйлов Євген Сергійович,

в.о. завідувача кафедри теоретичної та прикладної інформатики Харківського національного університету імені В.Н. Каразіна, Україна, Харків

Парфенюк Юрій Леонідович,

доктор філософії, викладач кафедри теоретичної та прикладної інформатики Харківського національного університету імені В.Н. Каразіна, Україна, Харків

Прокопович Світлана Валеріївна,

кандидат економічних наук, доцент кафедри економічної кібернетики і системного аналізу Харківського національного економічного університету імені Семена Кузнеця, Україна, Харків

Стрілець Вікторія Євгенівна,

кандидат технічних наук, доцент кафедри теоретичної та прикладної системотехніки Харківського національного університету імені В.Н. Каразіна, Україна, Харків

Угрюмов Михайло Леонідович,

професор, доктор технічних наук, професор кафедри теоретичної та прикладної системотехніки Харківського національного університету імені В.Н. Каразіна, Україна, Харків

Флейчук Марія Ігорівна,

професор, доктор економічних наук, професор кафедри маркетингу Львівського національного університету ветеринарної медицини та біотехнологій імені Степана Ґжицького, Україна, Львів

Чаговець Любов Олексіївна,

кандидат економічних наук, доцент кафедри економічної кібернетики і системного аналізу Харківського національного економічного університету імені Семена Кузнеця, Україна, Харків

Чумаченко Дмитро Ігорович,

доцент, кандидат технічних наук, доцент кафедри математичного моделювання та штучного інтелекту Національного аерокосмічного університету імені М. Є. Жуковського «Харківський авіаційний інститут», Україна, Харків

Шевченко Дмитро Олександрович,

аспірант кафедри теоретичної та прикладної системотехніки Харківського національного університету імені В.Н. Каразіна, Україна, Харків

Яковлев Сергій Всеволодович,

професор, доктор фізико-математичних наук, професор кафедри математичного моделювання та штучного інтелекту Національного аерокосмічного університету імені М. Є. Жуковського «Харківський авіаційний інститут», Україна, Харків

Hend Khalid Alkahtani,

Department of Information Systems of College of Computer and Information Sciences of Princess Nourah Bint Abdulrahman University, Saudi Arabia, Riyadh

Nitin Jain,

Professor, Department of Computer Science and Engineering, Chandigarh University, Punjab, India

Pradip M. Paithane,

Doctor of Philosophy, Assistant Professor, Department of Computer Engineering, Vidya Pratishthan’s Kamalnayan Bajaj Institute of Engineering and Technology, India, Pune

Ramalingappa Aparna,

Assistant Professor, Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bangalore, India

Sarita Jibhau Wagh,

Assistant Professor, Environment Science Department, T.C. College Baramati, India, Pune

Shilpa Chaudhari,

Associate Professor, Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bangalore, India

Suman Sansanwal,

Research Scholar, Department of Computer Science and Engineering, Chandigarh University, Punjab, India

Surender Singh Samant,

Doctor of Philosophy, Associate Professor, Department of Computer Science and Engineering, Graphic Era (Deemed to be University) Dehradun, India

Vedna Sharma,

Ph.D. Scholar, Department of Computer Science and Engineering, Graphic Era (Deemed to be University) Dehradun, India

Зміст журналу
«Системні дослідження та інформаційні технології»
за 2023 р.

ЗМІСТ № 1

Zgurovsky M., Yefremov K., Gapon S., Pyshnograiev I. Research of food security problems of the war-torn regions of Ukraine using geomatics methods 7

Vyklyuk Ya., Levytska S., Nevinskyi D., Hazdiuk K., Škoda M., Andrushko S., Palii M. Decision-tree and ensemble-based mortality risk models for hospitalized patients with COVID-19 23

Pankratova N.D., Musiienko D.I. Study of the underground tunnel planning. Cognitive modelling 37

Klymenko A.I., Podkolzin G.B. Modified seird model for describing of the COVID-19 epidemic 51

Kutsenko A.S., Kovalenko S.V., Kovalenko S.M. Generalization of the thermodynamic approach to multi-dimensional quasistatic processes 62

Pysarchuk O., Baran D., Mironov Yu., Pysarchuk I. Algorithms of statistical anomalies clearing for data science applications 78

Pradip M. Paithane, Sarita Jibhau Wagh, Sangeeta N. Kakarwal Optimization of route distance using K-NN algorithm for on-demand food delivery........ 85

Статкевич В.М. Модифікація мереж Петрі з антисипацією по позиції 102

Riznyk V.V. Researches and applications of the combinatorial configurations for innovative devices and process engineering 113

Бохонов Ю.Є. Інтегральні зображення додатно визначених ядер 122

Kovalov M. Approach to positional logic algebra 129

Zelensky K. Mathematical modelling of crystallization of polymer solutions 141

Відомості про авторів 151

ЗМІСТ № 2

Pankratova N.D., Grishyn K.D., Barilko V.E. Digital twins: stages of concept development, areas of use, prospects 7

Bidyuk P., Gavrilenko O., Myagkyi M. The algorithm for predicting the cryptocurrency rate taking into account the influence of posts of a group of famous people in social networks 22

Kulik A.S., Chukhray A.G., Havrylenko O.V. Information technology for creating intelligent computer programs for training in algorithmic tasks. Part 2: research and implementation 35

Zgurovsky M., Yefremov K., Gapon S., Pyshnograiev I. Assessment of the economical dimension of sustainable development of the Ukraine's regions based on the brightness of night lights 49

Кондратенко Н.Р., Снігур О.О., Кондратенко Р.М. Узагальнювальна інтервальна нечітка модель типу-2 для моніторингу станів складних систем з використанням експертних знань 63

Петренко А.І., Вохранов І.А. Нейронні мережі: дослідження правил прийняття ними рішень 74

Nakonechnyi O.G., Kudin G.I., Zinko P.M., Zinko T.P. Guaranteed root-mean-square estimates of the forecast of matrix observations under conditions of statistical uncertainty 86

Romanuke V.V., Romanov A.Y., Malaksiano M.O. A genetic algorithm improvement by tour constraint violation penalty discount for maritime cargo delivery 104

Курєннов С.С., Барахов К.П., Барахова Г.С. Застосування кривих Безьє для опису форми конструкції при оптимізації клейового з'єднання 127

Panibratov R.S., Bidyuk P.I. Estimation of the parameters of generalized linear models in the analysis of actuarial risks 139

Відомості про авторів 149

ЗМІСТ № 3

Zgurovsky M.Z. Academician Glushkov's legacy: human capital in the field of cybernetics, computing, and informatics at Igor Sikorsky Kyiv Polytechnic Institute 7

Dipa D. Dharmadhikari, Sharvari C. Tamane. Augmented security scheme for shared dynamic data with efficient lightweight elliptic curve cryptography 19

Havrylovych M.P., Danylov V.Y. Research on hybrid transformer-based autoencoders for user biometric verification 42

Zaychenko Yu., Zaichenko He., Kuzmenko O. Investigation of computational intelligence methods in forecasting at financial markets 54

Levenchuk L.B., Tymoshchuk O.L., Guskova V.H., Bidyuk P.I. Uncertainties in data processing, forecasting and decision making 66

Tharageswari K., Sundaram M., Santhosh R. A concatenation approach-based disease prediction model for sustainable health care system 81

Bukhori S., Verdy B.Y.N., Windi Eka Y.R., Januar A.P. Identification of lung disease types using Convolutional Neural Network and VGG-16 architecture 96

Lipianina-Honcharenko Khr., Bodyanskiy Ye., Sachenko A. Intelligent information system of the city's socioeconomic infrastructure 108

Matsuki Yo., Bidyuk P. Algorithm for simulating melting polar ice, Earth internal movement and volcano eruption with 3-dimensional inertia tensor 121

Melnyk I., Pochynok A. Basic algorithm for approximation of the boundary trajectory of short-focus electron beam using the root-polynomial functions of the fourth and fifth order 127

Відомості про авторів 149

ЗМІСТ № 4

Vedna Sharma, Surender Singh Samant. A multi-level decision-making framework for heart-related disease prediction and recommendation 7

Donets V.V., Strilets V.Y., Ugryumov M.L., Shevchenko D.O., Prokopovych S.V., Chagovets L.O. Methodology of the countries' economic development data analysis 21

Dorogyy Ya., Kolisnichenko V. Blockchain transaction analysis: a comprehensive review of applications, tasks and methods 37

Мартьянов Д., Виклюк Я., Флейчук М. Моделювання динаміки ринку криптовалют з використанням інструментів машинного навчання 54

Bazdyrev A. Semi-supervised inverted file index approach for approximate nearest neighbor search 69

Hend Khalid Alkahtani. Raising the information security awareness among social media users in the Middle East 76

Pradip M. Paithane, Sarita Jibhau Wagh. Novel modified kernel fuzzy c-means algorithm used for cotton leaf spot detection 85

Bazilevych K., Kyrylenko O., Parfeniuk Y., Yakovlev S., Krivtsov S., Meniailov I., Kuznietcova V., Chumachenko D. Information system for assessing the informativeness of an epidemic process features 100

Chaudhari S., Aparna R. Survey of image deduplication for cloud storage 113

Suman Sansanwal, Nitin Jain. A comprehensive survey on load balancing techniques for virtual machines 135

Відомості про авторів 148

Зміст журналу за 2023р. 150

Автори статей за 2023р. 152

АВТОРИ СТАТЕЙ ЗА 2023 р.

Андрушко Станіслав Дмитрович, №1
Баздирев Антон Андрійович, №4
Базілевич Ксенія Олексіївна, №4
Баран Данило Романович, №1
Барахов Костянтин Петрович, №2
Барахова Ганна Сергіївна, №2
Барілко Вєста Євгеніївна, №2
Бідюк Петро Іванович, №2,3
Бодянський Євгеній Володимирович, №3
Бохонов Юрій Євгенович, №1
Виклюк Ярослав Ігорович, №1,4
Вохранов Ілля Анатолійович, №2
Гавриленко Олена Валеріївна, №2
Гавриленко Олена Володимирівна, №2
Гаврилович Марія Павлівна, №3
Газдюк Катерина Петрівна, №1
Гапон Сергій Вікторович, №1,2
Грішин Костянтин Дмитрович, №2
Гуськова Віра Геннадіївна, №3
Данилов Валерій Якович, №3
Донець Володимир Віталійович, №4
Дорогий Ярослав Юрійович, №4
Єфремов Костянтин Вікторович, №1,2
Зайченко Олена Юріївна, №3
Зайченко Юрій Петрович, №3
Згуровський Михайло Захарович, №1,2,3
Зеленський Кирило Харитонович, №1
Зінько Петро Миколайович, №2
Зінько Тарас Петрович, №2
Кіріленко Олена Юріївна, №4
Колісніченко Вадим Юрійович, №4
Клименко Анастасія Іллівна, №1
Коваленко Світлана Миколаївна, №1
Коваленко Сергій Володимирович, №1
Ковальов Микола Олександрович, №1
Кондратенко Наталія Романівна, №2
Кондратенко Роман Михайлович, №2
Кривцов Сергій Олегович, №4
Кудін Григорій Іванович, №2
Кузнєцова Вікторія Олександрівна, №4
Кузьменко Олексій Віталійович, №3
Кулік Анатолій Степанович, №2
Курєннов Сергій Сергійович, №2
Куценко Олександр Сергійович, №1
Левенчук Людмила Борисівна, №3
Левицька Світлана Анатоліївна, №1
Ліп'яніна-Гончаренко Христина Володимирівна, №3
Малаксіано Микола Олександрович, №2
Мартьянов Дмитро Ігорович, №4
Мацукі Йошіо, №3
Мельник Ігор Віталійович, №3
Меняйлов Євген Сергійович, №4
Міронов Юрій Глібович, №1
Мусієнко Данило Ігорович, №1
Мягкий Михайло Юрійович, №2
Наконечний Олександр Григорович, №2
Невінський Денис Володимирович, №1
Палій Марина Анатоліївна, №1
Панібратов Роман Сергійович, №2
Панкратова Наталія Дмитрівна, №1,2
Парфенюк Юрій Леонідович, №4
Петренко Анатолій Іванович, №2
Писарчук Ілля Олексійович, №1
Писарчук Олексій Олександрович, №1
Пишнограєв Іван Олександрович, №1,2
Подколзін Гліб Борисович, №1
Починок Аліна Володимирівна, №3
Прокопович Світлана Валеріївна, №4
Різник Володимир Васильович, №1
Романов Андрій Юрійович, №2
Романюк Вадим Васильович, №2
Саченко Анатолій Олексійович, №3
Снігур Ольга Олексіївна, №2
Статкевич Віталій Михайлович, №1
Стрілець Вікторія Євгенівна, №4
Тимощук Оксана Леонідівна, №3
Угрюмов Михайло Леонідович, №4
Флейчук Марія Ігорівна, №4
Чаговець Любов Олексіївна, №4
Чумаченко Дмитро Ігорович, №4
Чухрай Андрій Григорович, №2
Шевченко Дмитро Олександрович, №4
Шкода Мирослав, №1
Яковлев Сергій Всеволодович, №4
Dipa D. Dharmadhikari, №3
Dr. N. Mohana Sundaram, №3
Dr. R. Santhosh, №3
Dr. Saiful Bukhori, №3
Dr. Sharvari Chandrashekhar Tamane, №3
Januar Adi Putra, №3
Hend Khalid Alkahtani, №4
K. Tharageswari, №3
Nitin Jain, №4
Pradip M. Paithane, №1,4
Ramalingappa Aparna, №4
Sangeeta N. Kakarwal, №1
Sarita Jibhau Wagh, №1,4
Shilpa Chaudhari, №4
Suman Sansanwal, №4
Surender Singh Samant, №4
Vedna Sharma, №4
Verdy Bangkit Yudho Negoro, №3
Windi Eka Yulia Retnani, №3