EXTENDED APPLIED DATA CLEANING METHODS IN OUTLIER DETECTION FOR RESIDENTIAL CONSUMER

Dacian I. **JURJ**, Dan D. **MICU**, Alexandru G. **BERCIU**, Mircea **LANCRAJAN**, Levente **CZUMBIL**, Andrei **BENDE**, Bogdan A. **MITRACHE**, Alexandru **MUREȘAN**

Technical University of Cluj-Napoca

dacian.jurj@ethm.utcluj.ro, Dan.Micu@ethm.utcluj.ro, Alexandru.Berciu@campus.utcluj.ro, lancranjanmircea98@gmail.com, Levente.CZUMBIL@ethm.utcluj.ro, andrei.bende28@gmail.com, bogdan.mitrache99@gmail.com, Alexandru.Muresan@ethm.utcluj.ro

Keywords: Energy measurement; Artificial intelligence; Statistics; Probabilistic computing; Data science; Data aggregation; Algorithms; Energy efficiency

Abstract*: This paper delves into the subject of outlier detection techniques tailored for unique datasets related to residential energy consumption. Building upon the current state of research [1] we introduce the Grubbs and Z-score methods and investigate a range of outlier detection strategies encompassing statistical, probabilistic, and machine learning algorithms. The findings underscore the importance of outlier detection in the Romanian residential energy sector.*

1. INTRODUCTION

Outlier detection in energy data plays a pivotal role in ensuring the accuracy and reliability of energy management systems [2]. By identifying and addressing anomalies, utilities and energy managers can gain a clearer understanding of consumption patterns, optimize energy distribution, and prevent potential system failures [3]. Moreover, detecting outliers, aids in eliminating data errors, facilitating more precise forecasting [4], and enhancing the overall efficiency of energy systems. In essence, it serves as a foundational step in refining energy data analysis and driving informed decision-making in the energy sector [5]. In the current study, serving as an extension to the article "Applied data cleaning methods in outlier detection for residential consumer" [1], the outlier detection methodologies tailored for specialized datasets related to residential energy consumption are examined. This exhaustive research covers a broad range of outlier detection techniques in data cleaning, from statistical and probabilistic methods to advanced machine learning algorithms. Notably, the Z-score [6] and Grubbs [7] methods are introduced and integrated into this expanded investigation, enhancing the analytical depth.

2. CONTEXT

Understanding the nature of a dataset is essential when applying algorithms or mathematical methods [8], especially when distinguishing energy consumption patterns across residential, industrial, and public buildings. Energy consumption characteristics differ significantly among residential homes, public infrastructures, and industrial facilities due to diverse usage patterns, energy needs, and operational demands [9]. Residential energy use is shaped by factors like home size, design, occupancy, construction materials, lifestyle, and appliance usage. Typically, residential energy patterns show spikes during morning and evening, reflecting daily routines, and dip during the night. Seasonal changes, such as increased heating or cooling needs during harsh weather, also play a role. Moreover, individual behaviors, household income, and occupants' education levels further influence these patterns [10]. Public structures, like schools, hospitals, and government offices, have energy patterns distinct from residential settings. Given their high occupancy and continuous operations, their energy use remains relatively consistent throughout the day, with higher consumption on weekdays than weekends. Unlike residences, seasonal fluctuations in energy use in public buildings are less pronounced, mainly due to their climate control systems [11].

Industrial sectors, including factories and warehouses, have unique energy consumption patterns driven by their production activities, machinery operations, and specific requirements. The energy demand in these settings is often high and consistent, influenced by machinery and equipment operations. However, energy use can vary based on production timelines, work shifts, and the nature of the industry. For instance, chemical industries might have different energy needs compared to textile ones. Industries linked to agriculture might see seasonal shifts in energy use, reflecting harvest times or changing product demands [12].

The importance of outlier detection in residential energy consumption arises from the diverse and ever-changing monthly energy use patterns. In contrast to public or industrial settings, monthly household energy consumption is influenced by numerous, primarily static, technical factors. These include home features, resident lifestyles, and energy use behaviors, which are often affected by seasonal changes [9].

In Romania, the energy sector is regulated by European directives mandating the installation of smart meters in residences by Distribution System Operators (DSO) [13]**Error! Reference source not found.**. However, the rollout of this initiative has been delayed, with DSOs typically manually reading meters every 3-6 months. While new regulations advocate for monthly readings, a disconnect exists between governmental intentions and the present capabilities of DSOs.

A significant shift occurred in the Romanian energy market in 2021 when it embraced free-market principles [14]. This change has spurred discussions on introducing various services and concepts, like demand response, differential tariffs, and local energy communities. Technically, these services necessitate precise monthly energy consumption data across all sectors, including households. Yet, due to current metering constraints, such data remains elusive.

For instance, the idea behind energy communities revolves around consolidating numerous households within similar geographic regions into a single organizational structure. This encourages them to actively engage in the energy market, promoting bi-directional energy exchanges (prosumers). In this context, precise monthly energy consumption predictions are crucial for securing smart energy contracts and negotiating favorable future [15].

A significant hurdle in Romania is obtaining accurate residential energy consumption data, primarily sourced from physical or digital energy bills. Given that current Romanian regulations require DSOs to check meters every 3-6 months, utility companies often resort to estimations. This approach frequently leads to substantial variations and anomalies in monthly billing, complicating the task of algorithms aiming to identify and rectify such patterns [16]. Consequently, due to these metering challenges, billed energy often doesn't mirror actual consumption, highlighting the need for a precise data adjustment method.

Such inconsistencies often manifest as outliers or anomalous data points, which significantly deviate from typical energy consumption trends. Identifying and addressing these outliers in the residential sector is pivotal for ensuring data accuracy in energy analysis, billing, and forecasting. This sets the stage for a robust data processing approach tailored [17] for the evolving needs of the Romanian energy market.

3. PREVIOUS RESULTS

The findings from previous research [1] indicated that the IQR method was the most effective, accurately identifying 69.45% of outliers. Both the MAD and MOVMAD methods displayed commendable results, detecting outliers at rates of 62.89% and 48.10%, respectively. In contrast, the LOF method's performance was less satisfactory, pinpointing just 23.83% of outliers.

Generally, the outlier detection techniques discussed in this study necessitate a substantial volume of data for precise outcomes. Despite the inherent challenges with residential energy consumption datasets, which typically comprise 12 data points annually, the researchers successfully adapted most of the algorithms. However, the parameters used for testing DBSCAN didn't match the outlier profiles, suggesting a need for further refinement of this method.

These insights underscore the effective implementation of the discussed methodology on datasets pertaining to the residential energy domain. Proper outlier identification is pivotal for data quality enhancement, deeper insights into energy consumption trends, and facilitating informed decisions in the residential arena. The findings advocate for the IQR method, succeeded by MAD and MOVMAD, as potential techniques for outlier detection in domestic energy data. Still, there's a call for more research to further hone these detection methods for superior precision.

4. METHOD USED

Given that statistical models have demonstrated greater efficiency with smaller data volumes in detecting outliers from the tested energy data consumption, the Z-score [18] and Grubbs [19] methods were selected for this study analysis. Both techniques are renowned for their efficacy in statistical analysis and outlier detection. By integrating these methods, the research aims to address the existing gaps and elevate the accuracy of data cleaning, ensuring more reliable and robust results in the realm of energy consumption analysis.

The Z-score [8], often referred to as the standard score, measures how many standard deviations a data point is from the mean of a set of data. It's a useful metric in statistics to identify outliers, as it quantifies the extent to which a particular observation deviates from the norm.

Mathematically, the Z-score for an individual data point x is calculated as:

$$\mathrm{Z} = \frac{x-\mu}{\sigma} \tag{1}$$

where:

- x is the individual data point.
- μ is the mean of the dataset.
- σ is the standard deviation of the dataset.

A Z-score [8] of 0 indicates that the data point's score is identical to the mean score. A Z-score of 1.0 indicates a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean. In many contexts, a Z-score above 2.0 or below -2.0 is considered an outlier, but this threshold can vary based on the specific application or field of study.

Grubbs' Test [19] is a statistical test used to detect outliers in a univariate data set that follows an approximately normal distribution. The test works by comparing the absolute deviation of a suspected outlier from the sample mean to the sample standard deviation. If this ratio is sufficiently large, the data point can be considered an outlier.
Mathematically, the Grubbs' statistic *G* for a given observation *xi* is calculated as:

$$G = \frac{\max\,(x_i - \bar{x})}{s} \tag{2}$$

where:

- x_i is the individual data point being tested.
- $\bar{x}$ is the sample mean.
- s is the sample standard deviation.

4. RESULTS

In this study, data was collected from 100 volunteering households. This data was then anonymized, cataloged, and subsequently analyzed. Energy consumption information was sourced from energy bills. However, it became evident that the recorded data didn't truly mirror the genuine energy usage patterns of the households. This discrepancy arose from the data processing system that relied on "estimations" and "energy meter readings". In Table 1 each data point represented the monthly energy consumption over the course of a year.

Table 1. Tested consumption energy data collected under a single database

User	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	74,00	79,00	74,00	74,00	61,00	66,00	65,00	54,00	54,00	17,00	127,00	58,00
2	NA	105,00	67,00	70,00	9,00	9,00	31,00	82,00	8,82	133,00	126,00	NA
3	68,00	76,00	68,00	72,00	648,00	95,00	252,00	116,00	54,00	351,00	100,00	128,00
4	131,00	60,00	99,00	88,00	52,00	58,00	73,00	52,00	62,00	51,00	53,00	66,00
5	76,00	74,00	67,00	53,00	128,00	NA	62,00	66,00	209,00	55,00	69,00	187,00
.	.	.	.	.	.	.	.	.	.	.	.	.
35	270,00	250,00	285,00	260,00	265,00	290,00	300,00	325,00	300,00	285,00	275,00	270,00
36	98,00	123,00	10,00	71,00	98,00	223,00	90,00	70,00	223,00	74,00	15,00	69,00
37	87,00	80,00	96,00	86,00	30,00	20,00	25,00	28,00	45,00	67,00	47,00	35,00
38	118,00	45,00	41,00	23,00	7,00	21,00	32,00	43,00	33,00	27,00	27,00	42,00
.	.	.	.	.	.	.	.	.	.	.	.	.
98	280,00	289,00	272,00	250,00	200,00	240,00	255,00	310,00	280,00	280,00	250,00	300,00
99	89,00	115,70	112,10	153,50	153,50	33,90	53,67	NA	137,10	118,00	NA	132,00
100	306,00	272,00	293,00	286,00	258,00	294,00	289,00	320,00	245,00	286,00	293,00	348,00

The proposed methods for detecting outliers, IQR (Interquartile Range), LOF (Local Outlier Factor), MAD (Median Absolute Deviation), and MOVMAD (Moving Median Absolute Deviation), underwent a rigorous validation process. This validation encompassed the analysis of energy consumption data derived from both public buildings [5] and residential buildings [1], reflecting their versatility and applicability across different contexts.

To gain a comprehensive understanding of these outlier detection methods, energy consumption data were collected and integrated into a unified database, as meticulously detailed in Table 2 and 3. Subsequently, a comparative analysis was conducted, pitting these techniques against the Z-score and Grubbs methods.

Prior to conducting the algorithm tests, any instances of missing data and irregularities were visually identified, allowing for an initial evaluation of the algorithms' accuracy. The abnormal data within the consumption profile was highlighted under the "HUMAN I" category and was specifically emphasized in Table 2 and 3 for User 5 and User 36. Likewise, the absence of a data point for User 5 and the visually identified outliers for User 36 were brought to attention in *figure 1* and *figure 2*.

Table 2. Algorithms precision over consumption energy of User 5

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
User 5	76,00	74,00	67,00	53,00	128,00	NA	62,00	66,00	209,00	55,00	69,00	187,00
HUMAN I						🟨			🟨			🟨
IQR	F	F	F	F	F	F	F	F	T	F	F	T
MAD	F	F	F	F	T	T	F	F	T	F	F	T
MOVMAD	F	F	F	F	F	F	F	F	T	F	F	F
LOF	F	F	F	F	T	T	F	F	T	F	F	T
Z-SCORE	F	F	F	F	F	F	F	F	T	F	F	F
GRUBBS	F	F	F	F	T	T	T	F	T	F	F	T

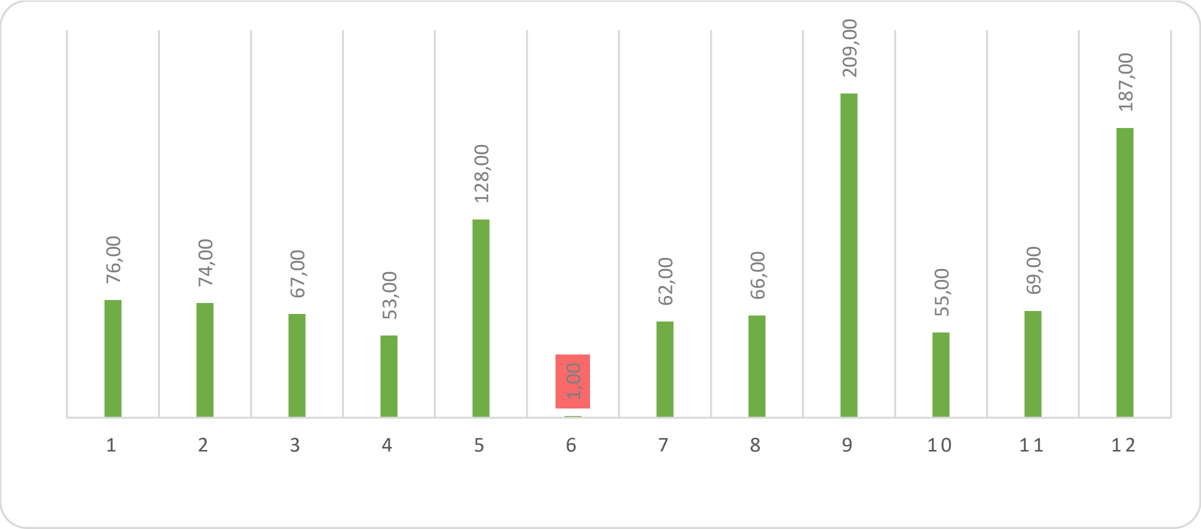

Fig. 1. User 5 missing data point for month 6

In order to evaluate the accuracy of the outlier detection methods, a scoring system was introduced. The scoring method was designed to avoid binary outcomes where points are awarded solely for 100% accuracy. The algorithms were classified as "T" (True) if they correctly identified the visually identified outlier and as "F" (False) if they either misclassified the outlier or highlighted a different one. This approach enables a more precise assessment by considering the granularity of points assigned, thereby reducing the presence of black and white scenarios.

If a method successfully identified all outliers, it was assigned a score of 1, as shown in previous study [1]. However, if the method erroneously classified non-outliers as outliers or failed to detect genuine outliers, the final score was calculated using the following formula:

$$\boldsymbol{Score} = \frac{\boldsymbol{Number\ of\ correct\ detections}}{\boldsymbol{Number\ of\ total\ detections}} \tag{3}$$

Table 3. Algorithms precision over consumption energy of User 36

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
User 36	98,00	123,00	10,00	71,00	98,00	223,00	90,00	70,00	223,00	74,00	15,00	69,00
HUMAN I						🟨			🟨			
IQR	F	F	F	F	F	T	F	F	T	F	F	F
MAD	F	F	T	F	F	T	F	F	T	F	F	F
MOVMAD	F	F	F	F	F	T	F	F	T	F	T	F
LOF	F	F	T	F	F	T	F	F	T	F	T	F
Z-SCORE	F	F	F	F	F	T	F	F	T	F	F	F
GRUBBS	F	F	F	F	F	F	F	F	F	F	F	F

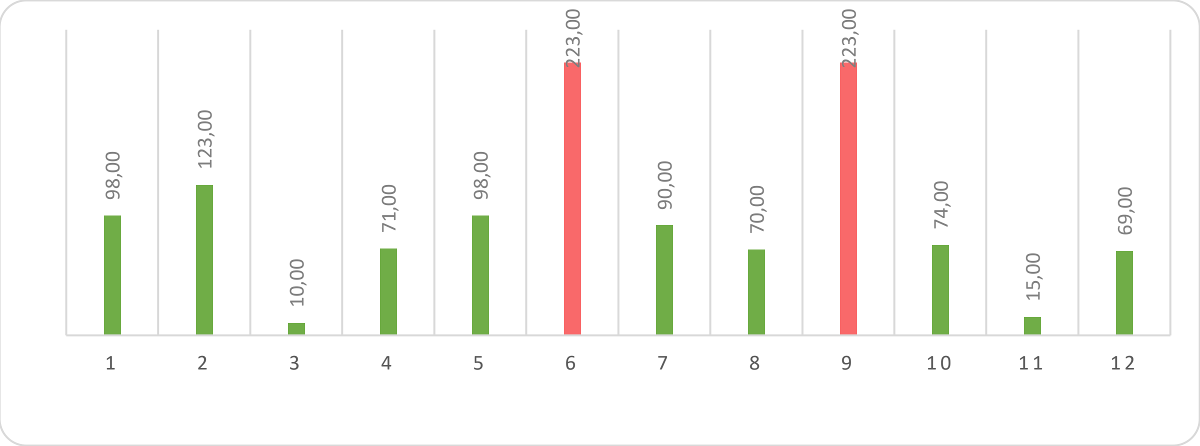

Fig. 2. User 36 extreme outlier data for month 6 and 9

Considering the results obtained from both Table 2 and Table 3, we can observe the following scores for the tested algorithms:

- **IQR** consistently performs exceptionally well in both evaluations, earning a perfect score of 1 in both instances.

- **MAD** demonstrates strong and consistent performance, achieving a score of 0.75 in Table 2 and a perfect score of 1 in Table 3.
- **MOVMAD** maintains a moderate level of accuracy, with scores of 0.33 in Table 2 and 0.66 in Table 3.
- **LOF** and **Z-score** exhibit similar performance, scoring 0.75 in Table 2 and 0.5 and 1, respectively, in Table 3.
- **Grubbs** shows variability in performance, with a score of 0.6 in Table 2 and a score of 0 in Table 3, suggesting room for improvement.

Table 4. The scoring applied to algorithm accuracy of detecting outliers

	IQR	MAD	MOVMAD	LOF	Z-SCORE	GRUBBS
User 1	1	1	0	0,5	1	0,5
User 2	0	0	0	0,22	0	1
User 3	0,5	1	0,5	0,5	0,5	0,66
User 4	1	1	0	0,33	1	1
User 5	1	0,75	0,33	0,75	0,33	0,6
.	.	.	.	.	.	.
User 35	1	1	0	0	0	0
User 36	1	1	0,66	0,5	1	0
User 37	1	1	1	0	1	1
User 38	1	1	0	0,25	1	0,5
.	.	.	.	.	.	.
User 98	1	1	0	0	0	1
User 99	0	0,75	0,5	0,75	0,33	0
User 100	1	1	1	0	0	0

Table 5. Final scoring and outlier detection accuracy over the tested algorithms:

	Scor	**%**
IQR	49,31	69,45
MAD	44,65	62,89
MOVMAD	34,15	48,10
LOF	16,92	23,83
Z-SCORE	43,14	60,76
GRUBBS	19,15	26,97

Applying the scoring method to the analyzed data from 100 users, as presented in Table 5, IQR emerges as the leading performer, securing a score of 49.31, which translates to an impressive accuracy rate of 69.45%. MAD also demonstrates robust performance, garnering a score of 44.65 and achieving an accuracy rate of 62.89%. MOVMAD maintains a reasonable accuracy rate of 48.10% while scoring 34.15. Z-score, while displaying promise with a score of 43.14, attains a slightly lower accuracy rate of 60.76%. Conversely, LOF and Grubbs exhibit lower scores and accuracy rates, with LOF registering 16.92 (23.83%) and Grubbs scoring 19.15 (26.97%). In summary, the IQR and MAD methods excel in accurately identifying outliers for this dataset, closely followed by MOVMAD and Z-score. Although LOF and Grubbs have their merits, they demonstrate relatively lower accuracy rates. Thus, for this particular dataset and evaluation, the IQR and MAD methods appear to offer the most reliable options for outlier detection.

5. CONCLUSION

When we compare the Z-score and Grubbs methods to the other outlier detection techniques, some interesting observations come to light: Firstly, both the Z-score and Grubbs methods hold their own in terms of their performance, falling somewhere in the middle of the pack. While they may not achieve the highest accuracy rates seen in some of the other methods, they display competitive scores and show potential for effectively identifying outliers in the context of residential energy consumption data. The IQR and MAD methods achieve the highest accuracy rates still Z-score perform significantly better than LOF and Grubbs. In scenarios where a balanced approach is essential, Z-score and Grubbs may be preferred choices. There is room for improvement in the Z-score and Grubbs methods. Further optimization could enhance their accuracy and reliability. In conclusion, the evaluation of the Z-score and Grubbs methods in this study suggests their potential as valuable tools for outlier detection. However, to further validate their effectiveness and robustness, future research endeavors should aim to test these methods with larger volumes of data. The scalability and adaptability of Z-score and Grubbs to more extensive datasets are crucial aspects that warrant exploration to ensure their reliability in various real-world applications and scenarios.

ACKNOWLEDGEMENT

This paper was financially supported by the Project "Network of excellence in applied research and innovation for doctoral and postdoctoral programs" / InoHubDoc, project co-funded by the European Social Fund financing agreement no. POCU/993/6/13/153437.

REFERENCES

[1] Jurj, Dacian I., Alexandru G. Berciu, Alexandru Muresan, Mircea Lancranjan, Levente Czumbil, Dan D. Micu, Andrei Bende, and Bogdan A. Mitrache. "Applied data cleaning methods in outlier detection for residential consumer." In *2023 10th International Conference on Modern Power Systems (MPS)*, pp. 01-04. IEEE, 2023.

[2] D. I. Jurj, D. D. Micu, L. Czumbil, A. G. Berciu, M. Lancrajan and D. M. Bărar, "Analysis of Data Cleaning Techniques for Electrical Energy Consumption of a Public Building," *2020 55th International Universities Power Engineering Conference (UPEC),* Turin, Italy, 2020, pp. 1-6, doi: 10.1109/UPEC49904.2020.9209781.

[3] D. Jurj, A. Polycarpou, L. Czumbil, A. Berciu, M. Lancranjan, D. Barar and D. Micu, "Extended

Analysis of Data Cleaning for Electrical Energy Consumption Data of Public Buildings," in 12th Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MEDPOWER), Online, 2020.

[4] D. Jurj, L. Czumbil, B. Bârgăuan, A. Ceclan, A. Polycarpou and D. Micu, "Custom Outlier Detection for Electrical Energy Consumption Data Applied in Case of Demand Response in Block of Buildings," Sensors, no. 21, p. 2946, 2021.

[5] Zhang, J., Zhang, H., Ding, S., & Zhang, X. (2021). Power Consumption Predicting and Anomaly Detection Based on Transformer and K-Means. *Frontiers in Energy Research*, *9*, 779587.

[6] V. Aggarwal, V. Gupta, P. Singh, K. Sharma and N. Sharma, "Detection of Spatial Outlier by Using Improved Z-Score Test," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 788-790, doi: 10.1109/ICOEI.2019.8862582.

[7] X. Guangcheng, C. Wenli, L. Xingzhi, Z. Ke, Z. Bo and S. Hongliang, "Research and Application of Verification Error Data Processing of Electricity Meter Based on Grubbs Criterion," 2019 International Conference on Smart Grid and Electrical Automation (ICSGEA), Xiangtan, China, 2019, pp. 13-17, doi: 10.1109/ICSGEA.2019.00012.

[8] S. K. Aggarwal, L. M. Saini and A. Kumar, "Electricity price forecasting in deregulated markets: A review and evaluation," Electrical Power and Energy Systems, no. 31, pp. 13-22, 2009.

[9] L. Perez-Lombard, J. Ortiz and C. Pout, "A review on buildings energy consumption information," Energy and Buildings, no. 40, pp. 394-398, 2008.

[10] Nikolaos Zografakis, Angeliki N. Menegaki, Konstantinos P. Tsagarakis, Effective education for energy efficiency, Energy Policy, Volume 36, Issue 8, 2008,

[11] Yang Wang, Jens Kuckelkorn, Fu-Yun Zhao, Di Liu, Alexander Kirschbaum, Jun-Liang Zhang, Evaluation on classroom thermal comfort and energy performance of passive school building by optimizing HVAC control systems, Building and Environment, Volume 89, 2015

[12] "Manufacturing Energy Consumption Survey 2018," Administration, U.S. Energy Information, Feb. 2021.

[13] "Smart Metering deployment in the European Union," [Online]. Available: https://ses.jrc.ec.europa.eu/smart-metering-deploymenteuropean- union.

[14] "Art 20 Energia electrică | Lege 123/2012," [Online]. Available: https://lege5.ro/Gratuit/gmzdenjwga/art-20-energia-electrica-lege-123-2012?dp=gyytmmbzge3dm.

[15] S. K. Aggarwal, L. M. Saini and A. Kumar, "Electricity price forecasting in deregulated markets: A review and evaluation," Electrical Power and Energy Systems, no. 31, pp. 13-22, 2009.

[16] Chanda, S.S., Banerjee, D.N. Omission and commission errors underlying AI failures. *AI & Soc* (2022).

[17] A. G. Berciu, D. Jurj, L. Czumbil, D. D. Micu and E. H. Dulf, "Energy Pulse – the Efficient Solution for Monitoring Electricity Consumption from Decentralized Data Sets," *2021 9th International Conference on Modern Power Systems (MPS)*, Cluj-Napoca, Romania, 2021, pp.

1-6, doi: 10.1109/MPS52805.2021.9492626.

[18] Vikas Khare, Cheshta Khare, Savita Nema, Prashant Baredar, Chapter 2 - Data visualization and descriptive statistics of solar energy system, Editor(s): Vikas Khare, Cheshta Khare, Savita Nema, Prashant Baredar, Decision Science and Operations Management of Solar Energy Systems, Academic Press, 2023,

[19] K. K. L. B. Adikaram, M. A. Hussein, M. Effenberger, T. Becker, "Data Transformation Technique to Improve the Outlier Detection Power of Grubbs' Test for Data Expected to Follow Linear Relation", *Journal of Applied Mathematics*, vol. 2015, Article ID 708948, 9 pages, 2015. https://doi.org/10.1155/2015/708948

ANALYSIS OF DIFFERENT DEEP LEARNING APPROACHES BASED ON DEEP NEURAL NETWORKS FOR PERSON RE-IDENTIFICATION

Adnan **RAMAKIĆ**[1], Zlatko **BUNDALO**[2], Dušanka **BUNDALO**[3]

[1] *Technical Faculty, University of Bihać, Bihać, Bosnia and Herzegovina,* [2] *Faculty of Electrical Engineering, University of Banja Luka, Banja Luka, Bosnia and Herzegovina,* [3] *Faculty of Philosophy, University of Banja Luka, Banja Luka, Bosnia and Herzegovina*

adnan.ramakic@unbi.ba, zlatbun2007@gmail.com, dusbun@gmail.com

Keywords: ***Deep Neural Network (DNN), Person Identification, Person Re-Identification, Convolutional Neural Network (CNN)***

Abstract: *In this work, different deep learning approaches based on deep neural networks for person re-identification were analyzed. Both identification and re-identification of people are frequently required in various fields of human life. Some of the most common applications are in various security systems where it is necessary to identify and track a particular person. In the case of person identification, the identity of a particular person needs to be established. In the case of re-identification, the main task is to match the identity of a particular person across different, non-overlapping cameras or even with the same camera at different times. In this work, three different deep neural networks were used for the purpose of person re-identification. Two of them were user-defined, while one of them is a pre-trained neural network adapted to work with a specific dataset. Two neural networks used were Convolutional Neural Networks (CNN). For the defined experiment, it was used own dataset with 13 subjects in gait.*

1. INTRODUCTION

Person identification and re-identification are important tasks in many aspects of human life. It is often necessary to determine the identity of a particular person, that is, to identify a particular person. This is a challenging task for which many methods have been developed in the last decades. Most of these methods are based on certain physiological or behavioral characteristics of the human body. The methods based on the mentioned characteristics are called biometric methods. Biometric methods are usually divided into two groups: physiological and behavioral biometric methods. Accordingly, there are methods based on a person's fingerprint or palm print, iris or retina (eye elements), face, gait, voice, or signature that are used in various applications.

The methods listed above are implemented in different ways and use different features based on the above characteristics. In general, an identification system can be divided into two parts. The first part is used to create a database (or in this paper denoted as *dataset*) in which images are usually captured for each person, e.g., using an RGB camera (Red, Green, Blue) or an RGB-D device (Red, Green, Blue - Depth). The images captured are stored in the database and depend on the method used. For example, if the implemented method is based on a person's face (face recognition), images containing people's faces are captured. On the other hand, if the method is based on gait (gait recognition), images with a person walking upright are usually captured and used. In the further course of the process, the aforementioned images can be subjected to different types of processing, depending on the method implemented. Features can also be extracted from the images and used, but this also depends on the method used. Accordingly, the extracted features may be also contained in the database. The second part of the identification system is the identification part, where the new image (or extracted features) of a particular person is matched with the images or features stored in the database. The above described can be roughly represented as in *figure* 1.

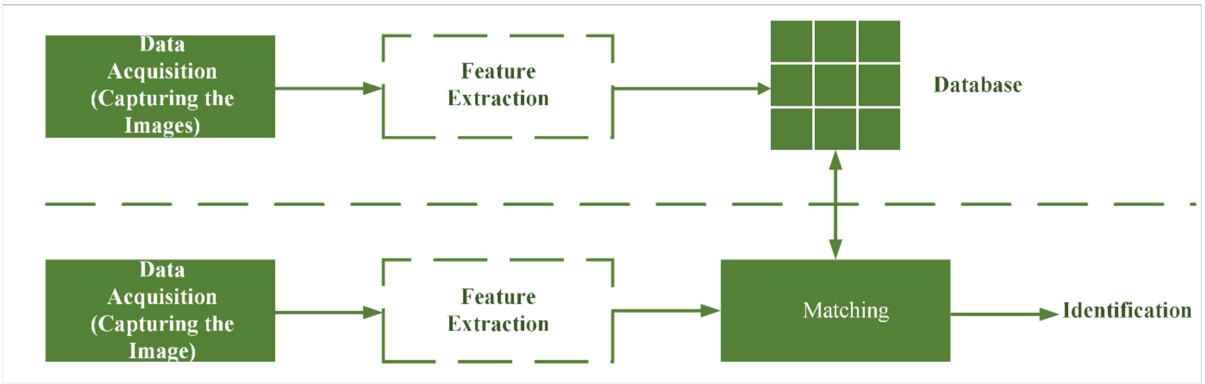
Data Acquisition (Capturing the Images)
Feature Extraction
Database
Data Acquisition (Capturing the Image)
Feature Extraction
Matching
Identification

Fig. 1. An Example of Identification System with Defined Steps

While person identification involves establishing the identity of a particular person, re-identification involves matching the identity of a particular person across different, non-overlapping cameras or even with the same camera at different time frames.

Nowadays, various methods for identification and re-identification are implemented using machine learning in such a way that a model is created, trained with the data, and the created model is then used for identification or re-identification tasks. Machine learning approaches typically use classifiers such as k-Nearest Neighbors (kNN) [1], Support Vector Machines (SVM) [2], or Linear Discriminant (LD) [3]. In addition to machine learning, deep learning approaches are also used, usually using a deep neural networks (DNN).

In this work, different approaches based on deep learning have been investigated. In this context, different deep neural networks were created and analyzed. An experiment was conducted with deep neural networks using a custom dataset. Accordingly, the experiment, the experimental setup and the obtained results have been described in the following chapters.

2. THE DATASET AND EXPERIMENTAL SETUP

In this work, a custom dataset was used for the defined experiment. The dataset used contains 13 people, during a walk (in gait), recorded with different camera positions. A stereo camera was used to create the dataset and multiple video footages were available for each person. The dataset was recorded in nice weather. The video footages have high resolution. Accordingly, the extracted images also have a high resolution. A drawback of the dataset can be during extraction of silhouettes, because some people wear clothes with similar colors as background. The size of the dataset (video footages in *.avi*) is about 1,5 GB. For each of the 13 people, different images were extracted from the recorded video footages and used in the experiment.

This was implemented so that in each video containing a particular person, the person was detected and tracked in each frame. To detect upright people, *vision.PeopleDetector* in Matlab [4] [5] may be used. In this context, a bounding box was formed around the person and this part of the scene was extracted and saved as an image in RGB format. This is shown in *figure* 2. The resolution of the extracted images containing only a person (green rectangular part in *figure* 2) was 185 x 375.

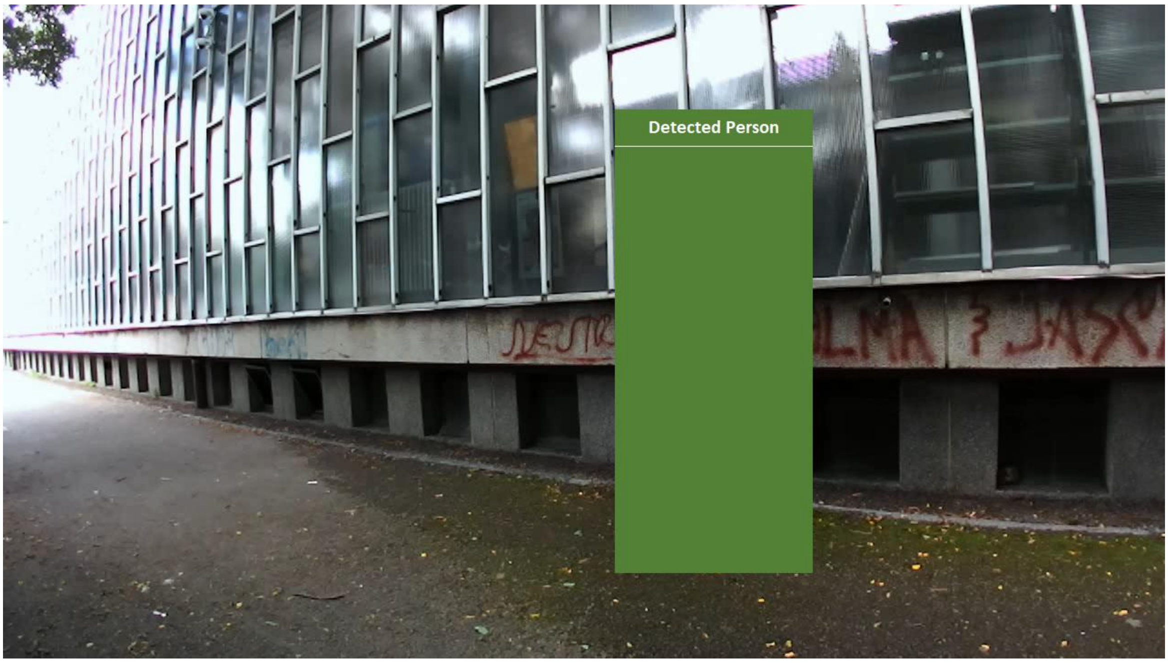

Fig. 2. Detected Person in One Video Frame

This procedure was performed for each of the 13 people. In other words, in the dataset there are 13 folders (*Person1*, *Person2* ... *Person13*) containing the images for each person. The mentioned procedure is illustrated in *figure* 3. The aforementioned extracted images are suitable for re-identification applications because the images are in RGB format and the people in all the captured images are wearing the same clothes. More specifically, said images can be used for short-term re-identification applications. For identification applications and long-term re-identification applications [6] [7] [8] [9] [10] [11], some longer-term features should be defined and used.

Various representations of silhouettes of people are often used as longer-term features, and many of the methods presented are based on them. An example of such a method is the well-known gait recognition method called Gait Energy Image (GEI) [12]. GEI is defined as an image containing silhouettes of a person over a gait cycle that are normalized, aligned, and temporally averaged. Some examples of silhouette images and GEI images from the Casia Dataset B [13] [14] [15] are shown in *figure 4* and *figure 5*.

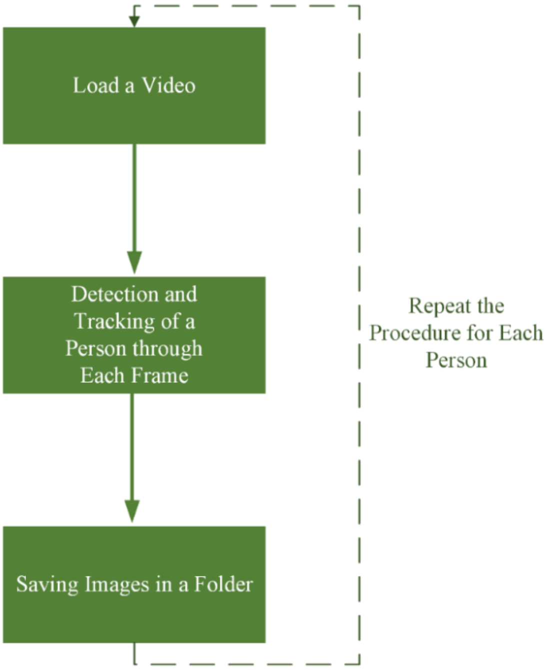
Load a Video
Detection and Tracking of a Person through Each Frame
Saving Images in a Folder
Repeat the Procedure for Each Person

Fig. 3. The Procedure for Dataset Creation

Once the dataset was created, three different deep neural networks were created and used for the experiment. The main idea was to create two different neural networks, with the first neural network having a feature input layer as the first layer. The features from the images would first be extracted and stored in a table and then used with the neural network created. In the case of the second neural network, the first layer is intended to be an image input layer. In this case, only images should be loaded to be used with the created neural network without prior feature extraction. In the third case, a pre-trained neural network was defined for use.

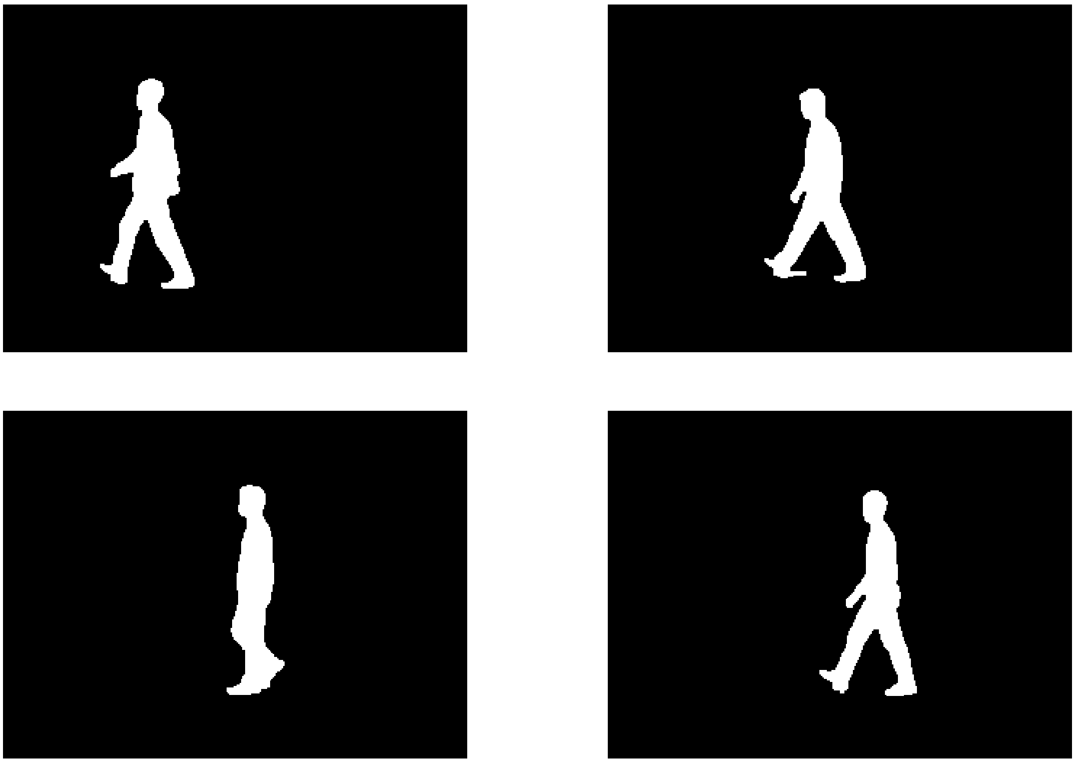

Fig. 4. Examples of Silhouette Images from Casia Dataset B [13] [14] [15]

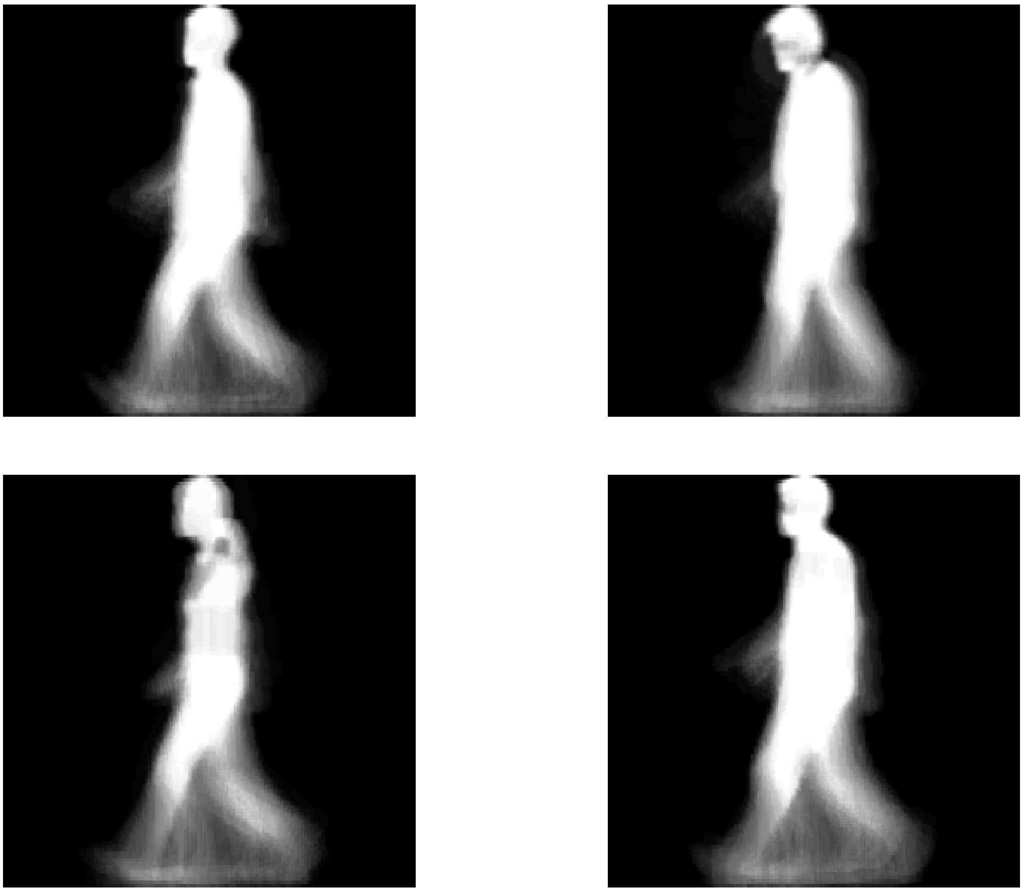

Fig. 5. Examples of GEI Images from Casia Dataset B [13] [14] [15]

Matlab was used for the mentioned experiment and for the creation of the dataset. For the creation of the dataset, a separate program was created for this purpose. It should be noted that Python, TensorFlow and Keras have also been analyzed, tested and used for the same purpose.

The first neural network (hereinafter marked with DNNf) uses extracted features from the images of the dataset. This was done using a *bag of visual words* (*bagOfFeatures* in Matlab, with defined parameters *VocabularySize - 500* and *PointSelection* as *Detector*) [16] [17], where the visual vocabulary was created by default from Speeded-Up Robust Features (SURF) [18]. The mentioned features were stored in a table. DNNf consists of seven layers, the first layer being the feature input layer (*featureInputLayer*). Different numbers and types of layers were tested, but with the mentioned seven layers and defined parameters, satisfactory results were obtained.

The seven defined layers are:

1. *featureInputLayer*
2. *fullyConnectedLayer*
3. *batchNormalizationLayer*
4. *reluLayer*
5. *fullyConnectedLayer*
6. *softmaxLayer*
7. *classificationLayer*.

The extracted features stored in the table were divided into a training and a testing part, with 70 percent used for training and 30 percent for testing. Other training options for the DNNf include 30 epochs, a learning rate of 0,001 and the Adaptive Moment Estimation Optimizer (Adam) [19] was used. The best results were obtained with the above settings.

The second neural network (hereinafter marked with DNNi) is a Convolutional Neural Network (CNN). The DNNi uses the images without prior feature extraction. The images were only loaded as the first layer is the image input layer (*imageInputLayer*). Also, in this case, different numbers and types of layers were analyzed and tested. With defined eight layers and defined parameters, satisfactory results were obtained.

DNNi consists of following eight layers:

1. *imageInputLayer*
2. *convolution2dLayer*
3. *batchNormalizationLayer*
4. *reluLayer*
5. *maxPooling2dLayer*
6. *fullyConnectedLayer*
7. *softmaxLayer*
8. *classificationLayer*.

The images used were also split in the ratio of 70 percent for training and 30 percent for testing. Other options defined for DNNi are 30 epochs, a learning rate of 0,001 and Stochastic Gradient Descent with Momentum (SGDM) [20] was used. Also, the best results were obtained with the above defined settings.

In addition to the deep neural networks created and described above (DNNf and DNNi), a pre-trained neural network was also used. The pre-trained neural network used is *GoogLeNet* [21] [22], a convolutional neural network. *GoogLeNet* [21] [22] was used and adopted to work with the dataset described above. This was done in such a way that two layers were replaced and adapted to the dataset. The layers mentioned are *fullyConnectedLayer* and *classificationLayer*. In the *fullyConnectedLayer*, the *OutputSize* parameter was set to *13*, which corresponds to the number of subjects in the dataset. The images used were split in the ratio of 70 percent for training and 30 percent for testing. Other training options for the *GoogLeNet* include 30 epochs, a learning rate of 0,001 and the SGDM was used, as in case DNNi.

3. RESULTS AND DISCUSSION

With the settings defined above and the neural networks described, the following results were obtained using the dataset described. In the case of DNNf, the accuracy was 90,8%. When DNNi was used, the accuracy was 91,7% which is higher compared to DNNf. In the case of *GoogLeNet*, pre-trained neural network, the accuracy was 99,4%. The results presented above are shown in table 1 and *figure 6*.

Table 1. The Obtained Results with Defined Settings and Used Dataset

Deep Neural Network Used	**Accuracy**
DNNf	90,8%
DNNi	91,7%
GoogLeNet	99,4%

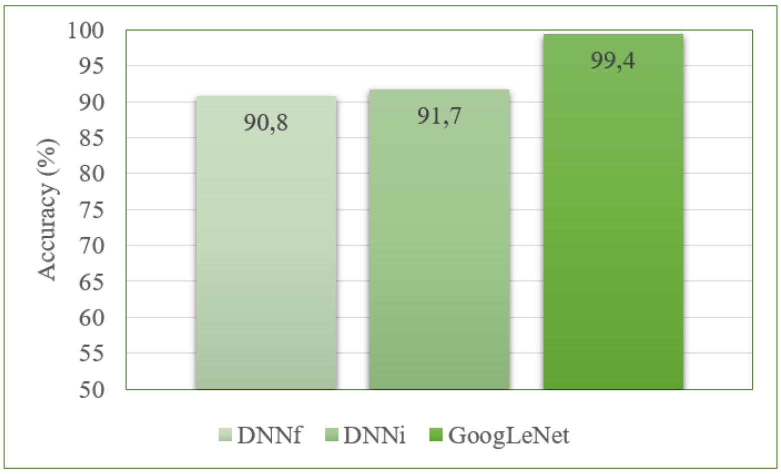

Fig. 6. The Obtained Results for the Deep Neural Networks Used

As can be seen from table 1 and *figure* 6, the pre-trained deep neural network *GoogLeNet* achieved the best overall result. This was to be expected, since *GoogLeNet* is a more complex neural network that has been pre-trained and validated on a large number of different images. With a relatively simple adaptation to use a custom dataset, the aforementioned deep neural network can easily be used for this type of application.

The created deep neural network, called DNNi, had the second best results and slightly better results compared to another created deep neural network (DNNf) that uses extracted features. On the other hand, DNNf has a much shorter training time. Moreover, the results of DNNf and DNNi can be improved by additional optimizations and adding some extra layers.

It should be noted that it is easier to work with deep neural networks such as DNNi and *GoogLeNet* compared to DNNf. The two deep neural networks mentioned, DNNi and *GoogLeNet,* have an image input layer as the first layer. This means that no explicit feature extraction is required in this case. For use with DNNi and *GoogLeNet*, only images containing people in gait should be loaded. In the case of DNNf, explicit feature extraction is required because the first layer is a feature input layer.

4. CONCLUSION

In this work, different deep learning approaches were analyzed. Person identification and re-identification applications are important in many areas of human life. In person identification, the identity of a particular person needs to be established. In person re-identification the main task is to match the identity of a particular person across different, non-overlapping cameras or with the same camera at different times. For example, in different security systems, some kind of identification or re-identification is often required.

Various methods have been developed for the aforementioned identification and re-identification applications. The mentioned methods are usually based on various physiological or behavioral characteristics of a person. Nowadays, identification and re-identification methods are usually implemented using various machine learning and deep learning approaches.

In this work, three different approaches based on deep neural networks were analyzed. For this purpose, two deep neural networks were created, while the third deep neural network used is pre-trained and adapted for use with a specific dataset. The first deep neural network created (DNNf) has a feature input layer as its first layer and uses extracted features from the images of the dataset. The second deep neural network (DNNi) is a convolutional neural network (CNN) and has as its first layer an image input layer into which only images to be used with said deep neural network are loaded. The third deep neural network used is the pre-trained neural network *GoogLeNet*.

The experiment with the defined deep neural networks was performed and the results were presented. For this purpose, a custom dataset containing 13 people in gait was used. The best overall result had the pre-trained deep neural network *GoogLeNet*.

In future research, it is planned to analyze and use a larger dataset containing a larger number of people in gait. In addition, it is also interesting to study different points of view and conditions where people wearing similar clothing. Accordingly, other deep neural network architectures will also be analyzed and studied.

REFERENCES

[1] L. E. Peterson, *K-Nearest Neighbor*, Scholarpedia, 4(2), 1883, 2009.

[2] S. R. Gunn, *Support Vector Machines for Classification and Regression,* ISIS Technical Report, 14(1), 5-16, 1998.

[3] R. A. Fisher, *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics 7(2), 179-188, 1936.

[4] N. Dalal and B. Triggs, *Histograms of Oriented Gradients for Human Detection*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 886-893, IEEE, 2005.

[5] Official Web Page of Mathworks, *vision.PeopleDetector (Documentation),* Link:*https://www.mathworks.com/help/vision/ref/vision.peopledetector-system-object.html* [Accessed 15/5/2023]

[6] K. Lenac, D. Sušanj, A. Ramakić and D. Pinčić, *Extending Appearance Based Gait Recognition with Depth Data*, Applied Sciences, 9(24), 5529, MDPI, 2019.

[7] A. Ramakić and Z. Bundalo, *Gait Recognition as an Approach for People Identification*, In: International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies, 717-726, Springer, 2023.

[8] A. Ramakić, Z. Bundalo and D. Bundalo, *An Example of Solution for Data Preparation Required for Some Purposes of People Identification or Re-Identification*, Journal of Circuits, Systems and Computers, *https://doi.org/10.1142/S0218126623501645*, World Scientific, 2022.

[9] A. Ramakić, Z. Bundalo and D. Bundalo, *A Method for Human Gait Recognition from Video Streams Using Silhouette, Height and Step Length,* Journal of Circuits, Systems and Computers, 29(7), 2050101, World Scientific, 2020.

[10] A. Ramakić, Z. Bundalo and Ž. Vidović, *Feature Extraction for Person Gait Recognition Applications*, Facta Universitatis, Series: Electronics and Energetics, 34(4), 557-567, 2021.

[11] A. Ramakić, D. Sušanj, K. Lenac and Z. Budalo, *Depth-based Real-time Gait Recognition*, Journal of Circuits, Systems and Computers, 29(16), 2050266, World Scientific, 2020.

[12] J. Han and B. Bhanu, *Individual Recognition Using Gait Energy Image*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(2), 316-322, IEEE, 2005.

[13] S. Yu, D. Tan, and T. Tan, *A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition*, In: 18th International Conference on Pattern Recognition (ICPR), 441-444, IEEE, 2006.

[14] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, *Robust View Transformation Model for Gait Recognition*, In: 18th International Conference on Image Processing, 2073-2076, IEEE, 2011.

[15] Official Web Page of the Institute of Automation, Chinese Academy of Sciences, Link: *http://www.cbsr.ia.ac.cn/english/Gait\%20Databases.asp* [Accessed 15/5/2023]

[16] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray, *Visual Categorization with Bags of Keypoints*, In: Workshop on Statistical Learning in Computer Vision (ECCV), 1-2, 2004.

[17] D. Nister and H. Stewenius, *Scalable Recognition with a Vocabulary Tree*, In: Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2161-2168, IEEE, 2006.

[18] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, *Speeded-Up Robust Features (SURF)*, Computer Vision and Image Understanding, 110(3), 346-359, Elsevier, 2008.

[19] D.P Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, arXiv preprint arXiv: 1412.6980, 2014.

[20] K.P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT PRESS, 2012.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, *Going Deeper With Convolutions*, IEEE Conference on Computer Vision and Pattern Recognition, 1-9, IEEE, 2015.

[22] Official Web Page of Mathworks, GoogLeNet, Link:*https://www.mathworks.com/help/deeplearning/ref/googlenet.html* [Accessed 15/5/2023]