BAJECE

e-mail: bajece@hotmail.com

Balkan Journal of Electrical & Computer Engineering

An International Peer Reviewed, Referred, Indexed and Open Access Journal

www.bajece.com

Vol : 10
No : 4
Year : 2022
ISSN : 2147 - 284X

It is abstracted and indexed in, Index Google Scholarship, the PSCR, Cross ref, DOAJ, Research Bible, Indian Open Access Journals (OAJ), Institutional Repositories (IR), J-Gate (Informatics India), Ulrich's, International Society of Universal Research in Sciences, DRJI, EyeSource, Cosmos Impact Factor, Cite Factor, SIS SIS Scientific Indexing Service, IIJIF, iiiFactor. ULAKBİM-TR Dizin.

Aim & Scope
The journal publishes original papers in the extensive field of Electrical-Electronics and Computer engineering. It accepts contributions which are fundamental for the development of electrical engineering, computer engineering and its applications, including overlaps to physics. Manuscripts on both theoretical and experimental work are welcome. Review articles and letters to the editors are also included.
Application areas include (but are not limited to): Electrical & Electronics Engineering, Computer Engineering, Software Engineering, Biomedical Engineering, Electrical Power Engineering, Control Engineering, Signal and Image Processing, Communications & Networking, Sensors, Actuators, Remote Sensing, Consumer Electronics, Fiber-Optics, Radar and Sonar Systems, Artificial Intelligence and its applications, Expert Systems, Medical Imaging, Biomedical Analysis and its applications, Computer Vision, Pattern Recognition, Robotics, Industrial Automation.

BAJECE
Balkan Journal of Electrical & Computer Engineering
An International Peer Reviewed, Indexed and Open Access Journal

ISSN: 2147- 284X
Vol: 10
No : 4
Year: October 2022

CONTENTS

BALKAN
JOURNAL OF
ELECTRICAL & COMPUTER ENGINEERING
(An International Peer Reviewed, Indexed and Open Access Journal**)**

Contact
Batman University
Department of Electrical-Electronics Engineering
Bati Raman Campus Batman-Turkey

Web: http://dergipark.gov.tr/bajece
https://www.bajece.com
e-mail: bajece@hotmail.com

Twisting Sliding Mode Control based Maximum Power Point Tracking

Korhan Kayisli and Ruhi Zafer Caglayan

Abstract— **For a world where energy demand is increasing day by day, available resources are constantly decreasing. At this point, it is very important to be able to benefit from the sun, which is the main energy source, with minimum damage to the environment. It is possible to produce electricity directly from sunlight through PV panels. Due to the limited efficiency of these panels, MPPT algorithms are always required. In this study, Sliding Mode Control (SMC) based Twisting Sliding Mode Control (T-SMC) MPPT, known for its robust structure, was performed and the results were compared with the classical SMC. The proposed MPPT algorithm is simulated with MATLAB/Simulink. The efficiency of T-SMC based MPPT is obtained as nearly 99%.**

Index Terms—**Boost Converter, Maximum Power Point Tracking, Photovoltaic Panel, Twisting Sliding Mode Control.**

I. INTRODUCTION

DURING TIME, humans were invented new things that got easier the human life. In the ancient ages, the wheel and the animal power were essential. After some eras, the first steam engine which need steam power was invented, then the jobs got easier. The steam has been produced by using fossil fuels in many years and at the following years. This steam (heat) energy has been used to turn the electric generators. With this production, the electricity has been started to use. The electric machines are more efficient than the fossil fuels, but it needed fossil fuel too due to this indirect generation. The environment has been damaged by the burning reaction of the fossil fuels and the burning reactions coproducts have been distorted in the Earth atmosphere and the greenhouse effect has been changed the climate. The main alternative energy sources named as renewable energy is emerged. More clean, sustainable, and efficient power is produced by using renewable energy. At this point, many researches have been carried out in order to benefit more from the sun, which is the main energy source of the world. Recently, photovoltaic (PV) panels have started to be produced in order to convert solar energy directly into electrical energy. Solar energy and PV panels are an important issue because they are renewable energy sources, the cost of sustainability is low, the variety of construction materials of PV panels, PV Panels can be produced for various power values. Due to the chemical properties of the materials used in PV panel production, very high efficiency cannot be achieved yet. One of the methods developed to obtain maximum efficiency from these panels is the Maximum Power Point Tracking (MPPT) technique. The aim is to examine new Maximum Power Point Tracking (MPPT) Method form using renewable sources that find more reliable, and efficient systems. There are many algorithms are used to obtain maximum power such as Incremental Conductance (IC), Open Circuit Voltage, Short Circuit Current, Perturb and Observe (P&O).

In a study, the recent application during the time period (1970s-1994), PV system configurations and the PV related issues were examined such as power conditions, protection, islanding, intermittent output and installation [1]. Another study analyzed the PV-thermal module of the real application in Spain. The main aim was to use solar energy to obtain electrical energy. Special type of PV panel was used in the application. From the analysis the radiation effect and the temperature effect were examined to the special PV-thermal panels. The collateral advantage of the system was heat regulation of the building [2] In [3], Hybrid PV/Thermal performance and system usagein the real live applications is researched. Different types of PV panels were used and optimized the system. Results show that the panel should be smaller than the collector unit. Bekker and Beukes were examined the optimal MPPT methods for examining the voltage control and voltage current control condition and they used Hill Climbing MPPT method. It is mentioned that the optimal control is current voltage control method on this study [4] Dachuan and Yuvarajan investigated the Hybrid PV and PEM Fuel Cell system and they tried to determine load sharing between the sources [5]. Dezso et.al. presented research that how to modelling a PV panel from using datasheet values. From known formulas and the given information on the datasheet, they used mathematical calculations and obtained the necessary, but not given data. [6]. In a study, series PV panels were connected different type DC/DC Converters such as boost converter and flyback converter to achieve more efficient systems with lower power stress. It is emphasized that the efficiency of flyback converter is better than boost converter [7]. A comparative MPPT study was performed to PV system and the performance of parasitic capacitance algorithm came to fore [8]. Another research was related with renewable energy sources and electrical vehicles. The main aim was to obtain a smart energy delivery from

KORHAN KAYISLI, is with Department of Electrical-Electronics Engineering Gazi University, Ankara, Turkey,(e-mail: korhankayisli@gmail.com).
https://orcid.org/0000-0001-8456-1478

RUHI ZAFER CAGLAYAN, is with Department of Electrical-Electronics Engineering Gazi University, Ankara, Turkey, (e-mail: ruhizaf08@gmail.com).
https://orcid.org/0000-0001-6585-7265

Manuscript received June 28, 2022; accepted July 25, 2022.
DOI: 10.17694/bajece.1137170

renewable energy sources to the loads [9]. Alan et.al. studied on assembly the renewable energy to obtain electrical energy sustainably introducing learning method [10]. Yamegueu et.al. examined the hybrid system without storage components. The results shown that if the load changes the PV panel validity effects the diesel generator's efficiency in the hybrid renewable and diesel generator systems [11]. Another study about to improve the efficiency of the series connected PV panels and they were shaded experimentally. The results shown that if the PV panels were directly connected to the DC/DC Converters, the system would work more efficiently. [12]. Sahoo et.al. tried to model of the PV systems by using Simscape simulation program and presented different type circuit topologies [13]. Palizban et.al. investigated the efficiency of the hybrid renewable system contains PV, wind, Fuel Cell, electrolyzer and super capacitor [14]. Active clamp interleved flyback converter was used to increase the efficiency of PV system and the converter was operated on DCM and CCM modes [15]. In another study, Z source inverter was used with PV panels. The DC output voltage was converted to AC by using this inverter driven with sinusoidal PWM. With this method, voltage gain was increased and voltage stresses were reduced [16]. A research was examined to find the maximum power point of the PV panel and it is aimed to obtain fast control with record the sun's radiation and temperature values on the lookup table by using the classical MPPT methods [17]. In a study, the power losses that occur as a result of the use of PV panels by connecting them in arrays and the formation of shadows was researched. It was tried to be solved by connecting each panel to the converter circuit one by one. For this purpose, buck-boost converter was also used to provide high efficiency [18]. A hybrid power system for low power electronics circuits and no-load requirement was investigated to obtain low-cost high efficiency structure [19]. It was aimed to supply the electrical energy needed by electric cars from renewable energy sources as solar energy and fuel cell. The performance was evaluated whether it charges the battery and provides the necessary energy [20]. Kale et.al. was explained to design of a highly efficient and reliable system called micro-inverter which fed from a solar system and to prevent load shedding problem, and islanding problem [21]. The researchers generally use single diode model for PV systems. While a research preferred to use multi-diode model and was simulated with MATLAB [22], others preferred to use the data given datasheets of manufacturers and generalized mathematical models [23]. In another study, PV and thermal panels were used together and artificial neural network were used to get the optimum power under changing temperature and irradiation conditions [24]. A new topology was presented by using buck boost and flyback topologies to provide low voltage to high voltage conversion using renewable energy sources such as PV panels. Benefits were emphasized such as reducing the number of circuit elements in this topology, reducing the diode and switch stress, highly efficiency, and recovering the lost energy [25]. A half wave converter was used to improve the efficiency of PV systems under variable temperature and irradiation. It was aimed to reduce the inductor size and obtain high efficiency with simulations [26]. In [27], Belkaid et.al were aimed to develop the P&O algorithm, also proposed a modified equivalent sliding mode MPPT for better dynamic behavior [28]. Alhammad et.al. was designed a system contains PV panels and thermoelectric generator to feed electrical vehicles efficiently [29]. In a study, PV/TEG hybrid system was presented and controlled with sliding mode control to cope with temperature and irradiation changes [30]. Raju and Mikkili was explored different connection types such as serial, parallel, serial-parallel, Honey-Comb under shading conditions [31].

Additionally, the effects of single diode PV model on current, voltage, resistances and ideality factor were investigated [32]. Mnati et.al. compared the efficiency and dynamic behaviors of P&O, IC, Constant voltage,Open Circuit Voltage techniques [33]. For fast response, fuzzy logic control was used for MPPT and obtained 99% efficiency [34]. Beyarslan was aimed to design a micro-grid structure by balancing the energy need with storage systems aswell as renewable energy sources such as wind, solar and hydroelectric systems. In this way, a system was designed to meet 100% of a region's energy needs [35]. From the literature review, sliding mode control (SMC), fuzzy logic control (FLC), artificial neural network (ANN) and some other methods are much popular to obtain better MPPT performances. Because the performance of traditional methods is limited and modified to achieve the maximum.

There are some studies have been performed to get maximum efficiency under variable atmospheric conditions such as System Identification based ARV MPPT [39], a voltage scanning-based MPPT [40], P&O and INC MPPT Methods using FPGA [41] and system identification-based MPPT [42].

In this study, it is aimed to obtain maximum efficiency from a PV system by using a robust control algorithm. SMC is well known method to control the system under parameter changes and distributions. The amount of radiation is not fixed and is variable. A robust control structure can be obtained with SMC under variable irradiation condition with better MPPT performance. In the meantime, second order SMC algorithms such as twisting SMC (T-SMC) appears with better dynamic behavior. A boost converter fed from PV panel is controlled with T-SMC algorithm to obtain more efficiency and robust structure.

II. PV Panel Model

PV panels are special devices that have been made Silicon and the other materials have p-n junction cells. If the sun light hit this p-n junction and the sun lights contained energy is enough to increase electrons power greater than the band gap (the energy between electrons orbital energy to free electron level). The electrons freed by effect of irradiation with the semiconductor structure that generates electron flow [22] This is the working principle of the PV Panel. In this part The PV Panel Design is detailly discussed [6],[22],[36].

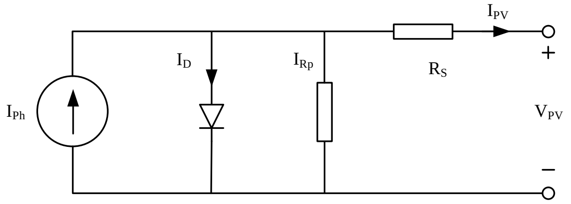


Fig. 1. PV Panel equivalent circuit model (single diode)

The traditional model of PV panel is single diode model and it contains one controlled current source, one diode (controlled current source too), one paralleled, one series resistance.

There are three current loop in this circuit;

- One of them is I_{Ph} current loop
- The second is between I_D and R_p current loop
- The third loop is the I_{PV} current loop

$$I_{PV} = I_{Ph} - I_D - I_{R_p} \tag{1}$$

The current of the diode is given Eq.2 [27,40].

$$I_D = I_{sat} * (e^{k*\frac{V_d}{T_K}} - 1) \tag{2}$$

By using Eq.2 in Eq.1, the final diode current is obtained as Eq.3.

$$I_D = I_{sat} * (e^{\frac{V_d}{n_{cell}*V_{thermal}}} - 1) \tag{3}$$

$$V_t = \frac{quality*k*T_{STC}}{q} \tag{4}$$

$$I_{Ph} = [I_{SCR} + k_i(T - T_{ref})](\frac{S}{S_{ref}}) \tag{5}$$

$$I_{PV} = I_{Ph} - I_{sat} * (e^{\frac{V_{PV}+I_{PV}*R_S}{n_{cell}*V_{thermal}}} - 1) - \frac{V_{PV}+I_{PV}*R_S}{R_p} \tag{6}$$

From above equations, the open circuit, short circuit and at Maximum Power Point expressions can be obtained as follows:

$$I_{short} = I_{Ph} - I_{sat} * (e^{\frac{I_{short}*R_S}{n_{cell}*V_{thermal}}} - 1) - \frac{I_{short}*R_S}{R_p} \tag{7}$$

$$0 = I_{Ph} - I_{sat} * (e^{\frac{V_{PV}}{n_{cell}*V_{thermal}}} - 1) - \frac{V_{PV}}{R_p} \tag{8}$$

$$I_{PV_{MPPT}} = I_{Ph} - I_{sat} * (e^{\frac{V_{PV_{MPPT}}+I_{PV_{MPPT}}*R_S}{n_{cell}*V_{thermal}}} - 1) - \frac{V_{PV_{MPPT}}+I_{PV_{MPPT}}*R_S}{R_p} \tag{9}$$

III. Boost Converter Model

The most commonly used dc-dc converter type in MPPT process is the boost converter. A classic boost converter circuit is shown in the Fig.2.

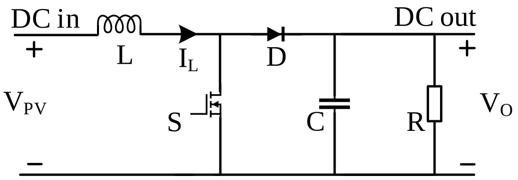


Fig. 2. Boost Converter

The circuit is examined in two cases, depending on whether the power switch S is on and off. Accordingly, the equivalent of the circuit is as in Fig.3. when the switch S is on [27].

- (Switch S on State):
  $0 \leq t \leq d * t_{switch}$
- (Switch S off State)
  $d * t_{switch} \leq t \leq t_{switch}$

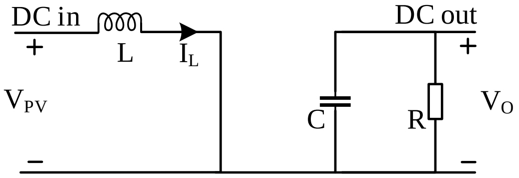


Fig. 3. Boost Converter (Switch S on State)

$$i_L = i_{L_{initial}} + \frac{1}{L}\int_0^t V_L(t)dt \tag{10}$$

Where,

$$V_L(t) = V_{PV} \tag{11}$$

$$i_{L_{initial}} = i_{L_{min}} \tag{12}$$

$$i_L = i_{L_{min}} + \frac{1}{L} * V_L * d * t_{switch} \tag{13}$$

$$\Delta i_L = \frac{V_L(t)*d*t_{switch}}{L} \tag{14}$$

In case the switch is off, the equivalent of the circuit is as given in Fig.4.

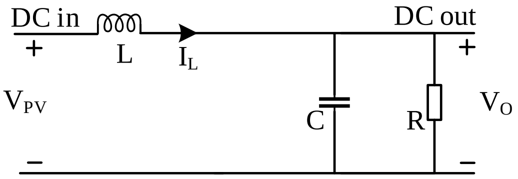


Fig. 4. Boost Converter (Switch S off State)

$$i_L = i'_{L_{initial}} + \frac{1}{L}\int_0^t V'_L(t)dt \tag{15}$$

Where,

$$V'_L(t) = V_{PV} - V_O \tag{16}$$

$$i'_{L_{initial}} = i_{L_{max}} \tag{17}$$

$$i_L = i_{L_{max}} - \frac{1}{L} * (V_O - V_{PV}) * (t_{switch} - d * t_{switch}) \tag{18}$$

$$\Delta i_L = \frac{(V_O-V_{PV})*(t_{switch}-d*t_{switch})}{L} \tag{19}$$

The relationship between output voltage and input voltage is

$$V_{PV} * t_{on} + (V_{PV} - V_O) * t_{off} = 0 \tag{20}$$

$$t_{on} = d * t_{switch} \tag{21}$$

$$t_{off} = t_{switch} - d * t_{switch} \tag{22}$$

$$V_{in} * d * t_{switch} = -(V_{in} - V_{out}) * (t_{switch} - d * t_{switch}) \tag{23}$$

$$\frac{V_O}{V_{PV}} = \frac{1}{1-d} \tag{24}$$

From these equations,

$$\Delta i_L = \frac{V_O*(d-d^2)*t_{switch}}{L} \tag{25}$$

In the on state the capacitor discharge and the output voltage is produced by capacitor. The capacitors voltage equation is:

$$V_O = V_{max} + \frac{1}{C} * (-I_o * d * t_{switch}) \tag{26}$$

$$\Delta V_O = \frac{I_O*d*t_{switch}}{C} \tag{27}$$

IV. Twisting Sliding Mode MPPT

Generally, algorithms such as P&O and IC are used for MPPT process. As an alternative to these methods, SMC-based MPPT can also be preferred. The SMC is a widely used method in nonlinear systems, known as robust controller. There are also higher order SMC structures available such as (T-SMC). In this study, T-SMC is designed for MPPT process.

The main difference between SMC and T-SMC is that the trajectories oscillate with twisting on the sliding surface instead of chattering. SMC consists of two basic steps: determining the sliding surface that will enable the boost converter circuit to work as desired and determining the control rule that directs the system to this sliding surface and enables it to operate on this surface. In order to perform the MPPT operation with the boost converter, firstly, the sliding surface can be defined as Eq.28. Here the sliding surface and its derivative are equal to zero. As long as the system trajectories reach and stay on the sliding surface, the system will operate at the maximum power point [43].

$$V(x,t) = \frac{\partial P_{PV}}{\partial V_{PV}} = V_{PV}\left(\frac{\partial I_{PV}}{\partial V_{PV}} + \frac{I_{PV}}{V_{PV}}\right) = 0 \tag{28}$$

The T-SMC contains two control laws named as switching and equivalent control. The equivalent control and T-SMC can be obtained by using the following expression.

$$x = \begin{bmatrix} I_L \\ V_O \end{bmatrix} \tag{29}$$

$$\dot{\mathrm{x}} = f(\mathrm{x}) + g(\mathrm{x})u \tag{30}$$

$$f(x) = \begin{bmatrix} \frac{V_{PV}-V_O}{L} \\ \frac{I_{PV}}{C_O} - \frac{V_O}{RC_O} \end{bmatrix} \quad g(x) = \begin{bmatrix} \frac{V_O}{L} \\ -\frac{I_{PV}}{C_O} \end{bmatrix} \tag{31}$$

$$\dot{V} = \left[\frac{\partial V}{\partial x}\right]^T \dot{V} = \left[\frac{\partial V}{\partial x}\right]^T \left(f(x) + g(x)u_{eq}\right) = 0$$

$$u_{eq} = \frac{\left[\frac{\partial V}{\partial x}\right]^T f(x)}{\left[\frac{\partial V}{\partial x}\right]^T g(x)} = 1 - \frac{V_{PV}}{V_O} \tag{32}$$

$$u_{T-SMC} = u_{eq} - \alpha_1 sign(V) - \alpha_2 sign(\dot{V}) \tag{33}$$

Lyapunov Stability Analysis

The basic logic of Lyapunov's theorem is that a continuously decreasing definite positive function must go to zero. If we can find the negative time derivative (*$\dot{L}(x)$*) of a strictly positive function *L(x)*, then the system is asymptotically stable.

$$L(x,t) = \frac{1}{2}\left(V(x,t)\right)^2 \tag{34}$$

$$\dot{L} = \left[\frac{\partial L}{\partial x}\right]^T \dot{x} = \left[\frac{\partial L}{\partial I_{PV}}\right]\left(-\frac{V_O}{L}(1-u) + \frac{V_{PV}}{L}\right) \tag{35}$$

$$\dot{V} = \left[\frac{1}{\mathrm{V_{PV}}} - \frac{\mathrm{I_{PV}}}{\mathrm{V_{PV}^2}}\frac{\partial \mathrm{V_{PV}}}{\partial \mathrm{I_{PV}}} + \frac{\mathrm{q}}{\mathrm{N_S}\eta \mathrm{V_T}}\frac{\partial \mathrm{I_{PV}}}{\partial \mathrm{V_{PV}}}\frac{\partial \mathrm{V_{PV}}}{\partial \mathrm{I_{PV}}}\right]\left(-\frac{V_O}{L}(1-u) + \frac{V_{PV}}{L}\right) \tag{36}$$

The first derivative of $\mathrm{I_{PV}}$ and $\mathrm{V_{PV}}$ are defined as in Eq.37 and Eq.38, respectively.

$$\frac{\partial \mathrm{I_{PV}}}{\partial \mathrm{V_{PV}}} = -\frac{\mathrm{q}}{\mathrm{N_S}\eta \mathrm{V_T}}\mathrm{I_D}\,\mathrm{ex\,p}\left(\frac{\mathrm{q}}{\mathrm{N_S}\eta \mathrm{V_T}}\right) < 0 \tag{37}$$

$$\frac{\partial \mathrm{V_{PV}}}{\partial \mathrm{I_{PV}}} = -\frac{\mathrm{N_S}\eta \mathrm{V_T}}{\mathrm{q}}\ln\left(\frac{\mathrm{I_D}}{\mathrm{I_{PH}}+\mathrm{I_D}-\mathrm{I_{PV}}}\right) < 0 \tag{38}$$

The sign of the first term (Eq.35) is positive when $\frac{\partial \mathrm{V}}{\partial \mathrm{I_{PV}}} > 0$.

$$\dot{x} = -\frac{V_O}{L}\left(1 - \left(1 - \frac{V_{PV}}{V_O}\right) - u_{T-SMC}\right) + \frac{V_{PV}}{L} \tag{39}$$

$$V\dot{V} = V\left[\frac{\partial V}{\partial I_{PV}}\right]\frac{V_O}{L}\left(-\alpha_1 sign(V) - \alpha_2 sign(\dot{V})\right) \tag{40}$$

V and *$\dot{V}$* always have different signs and Lyapunov stability criteria is ensured.

V. Simulation of Proposed System

The PV panel, boost converter and T-SMC based MPPT are simulated by using MATLAB/Simulink software. The simulation of the proposed system is shown in Fig.5.

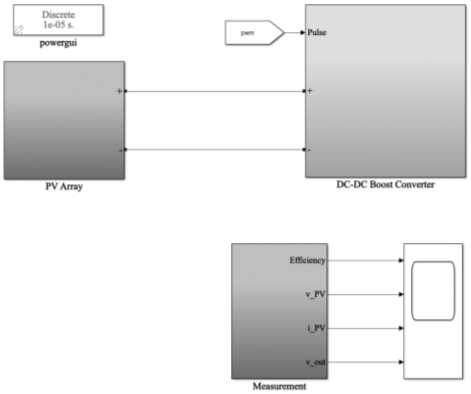


Figure 5. Simulation of proposed system

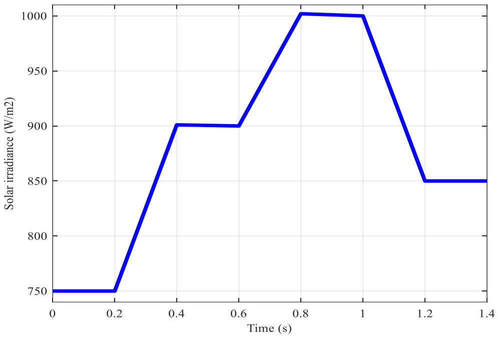


Figure 6. Solar irradiation profile

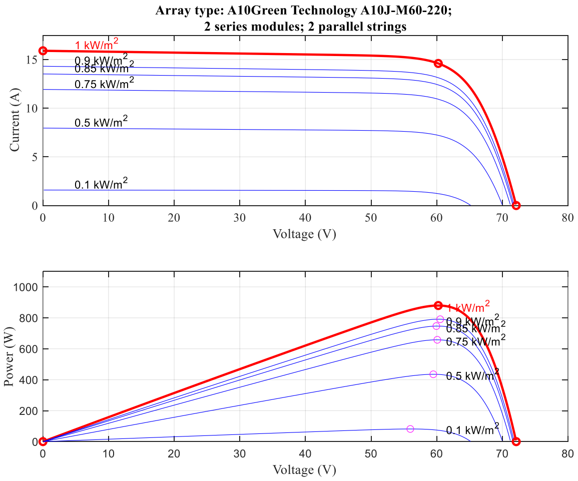


Figure 7. Characteristics of PV array

A variable solar irradiance profile is applied to the PV panel. 2 series and 2 parallel strings A10J-M60-220 solar panel is used on the proposed system. The solar irradiation profile and voltage-current, voltage-power characteristic of the PV array are shown in Fig.7.

The parameters of PV array, boost converter, SMC and T-SMC are given in Table.1.

TABLE 1. SIMULATION PARAMETERS

PARAMETERS OF PV ARRAY	
Parameter	**Value**
Maximum Power (W)	213.15
Cells per Module (N_{cell})	60
Open circuit voltage V_{OC} (V)	36.3
Short circuit current I_{SC} (A)	7.84A
Voltage at MPP V_{MP} (V)	29
Current at MPP I_{MP} (A)	7.35
Temperature (ºC)	25
Parallel strings (N_P)	2
Series connected modules per string (N_S)	2
PARAMETERS OF BOOST CONVERTER	
Parameter	**Value**
Inductor L (mH)	2.2
Capacitor C (μF)	100
Resistance load (Ω)	25
Switching frequency (kHz)	10
Diode forward voltage (V)	0.8
IGBT forward voltage (V)	1
CONTROLLER PARAMETERS	
Parameter	**Value**
SMC (λ)	0.0487
T-SMC ($\boldsymbol{\alpha}_1$)	0.715
T-SMC ($\boldsymbol{\alpha}_2$)	0.014

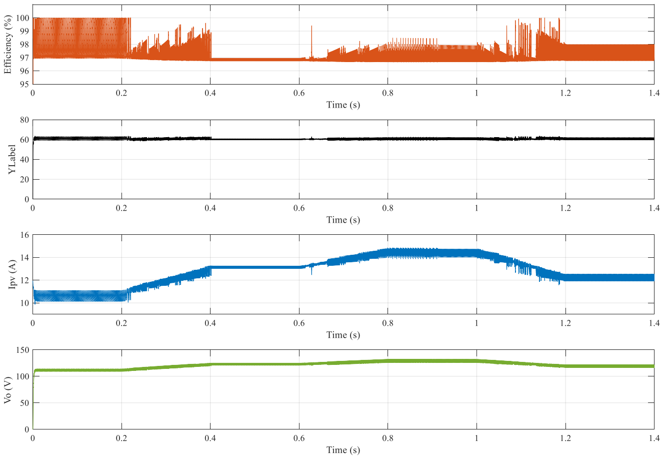


Figure 8. Simulation results of SMC based MPP

VI. RESULTS

There are two simulations are performed in this study. SMC based MPPT and T-SMC based MPPT algorithms are used to show and evaluate the performance of the proposed T-SMC based MPPT algorithm. The results are presented with efficiency, V_{PV}, I_{PV} and V_O of the boost converter. The results of SMC based MPPT and T-SMC based MPPT algorithms are shown in Fig.8. and Fig.9, respectively. The efficiency of SMC changes between 97%-98% for the high irradiation conditions. But this situation is improved as 97%-100% by using T-SMC. The efficiency of SMC and T-SMC based MPPT algorithms are shown in Fig.10. The efficiency performance of both algorithm is compared by presenting Fig.10. Also, the mean efficiency, maximum and minimum values are given in Table 2.

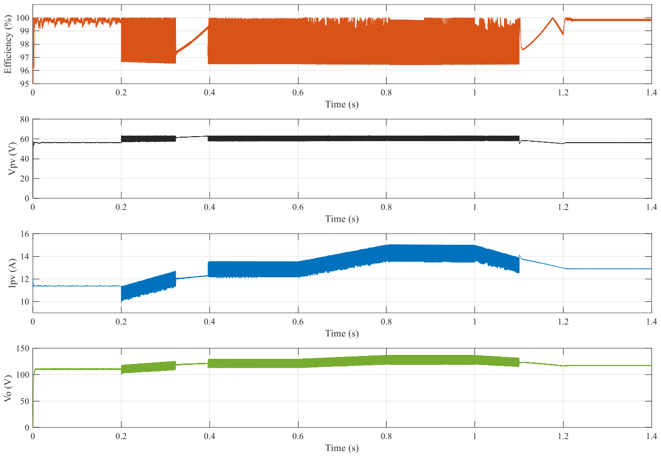


Figure 9. Simulation results of T-SMC based MPPT

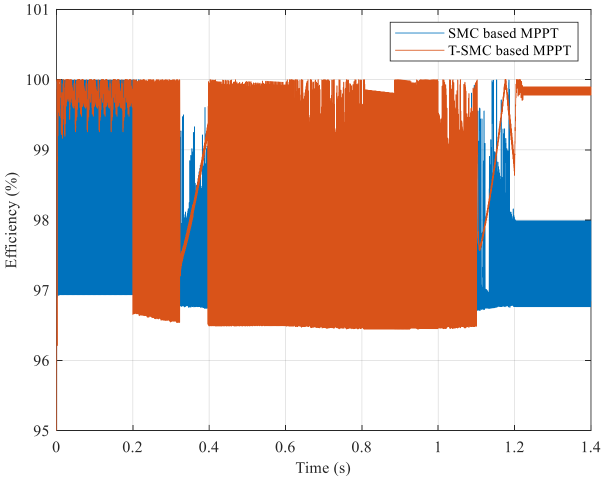


Figure 10. Efficiency of SMC and T-SMC based MPPT

TABLE 2. PERFORMANCE OF SMC AND T-SMC BASED MPPTS

	SMC MPPT	**T-SMC MPPT**
Efficiency (Eff.) (Mean)	97.1496%	98.8777%
Fluctuation on Eff. (Min-Max)	96.66%-100%	96.45%-100%
V_{PV} (Max)	62V	63V
I_{PV} (Max)	14.8A	15.02A
Fluctuation on V_O (Min-Max)	132.3V	136.8V

VII. CONCLUSION

In this study, T-SMC based MPPT algorithm is proposed and its performance is compared with the classical SMC based algorithm. The dynamic performance of T-SMC is better than SMC when the irradiation levels approach maximum value as

1000 W/m^2. This situation effects the mean efficiency of MPPT algorithms and T-SMC has 98.8777% efficiency. The efficiency performance is better than SMC based MPPT algorithm. Also, the proposed system cannot much be affected from the irradiation changes and the efficiency is generally higher than 97%. The dynamic response of the proposed MPPT technique is much sufficient especially irradiation changes. Also, mean efficiency is more than 98%. In future studies, efforts will be made to minimize the chattering problem and improve MPPT performance by adapting different control techniques and parameter optimization.

NOMENCLATURE

I_{SCR}	Short circuit current at reference temperature
K	Boltzmann constant
k_i	Cell's short circuit current temperature coefficient
n_{cell}	PV panel total cell number
q	electron charge
S	Sun irradiance
S_{ref}	Solar radiation reference
T	Cell temperature
T_{ref}	Cell's reference temperature
T_{STC}	Standard test condition temperature
V_d	Diodes terminals potential difference
V_L	Inductance voltage
$V_{thermal}$	Thermal voltage
d	duty cycle

REFERENCES

[1] R. Ramakumar, “Photovoltaic applications”, *IEEE Power Engineering Review*, pp. 17-21, April 2004.

[2] A. Lloret, J. Andreu, J. Merten, J. Puigdollers, O. Aceves, L. Sabata, M. Chantant, and U. Eicker, “Large grid-connected hybrid pv system integrated in a public building”, *Progress in Photovoltaics: Research and Applications Prog. Photovolt. Res. Appl.* , vol. 6, pp. 453-464, 1998.

[3] R. Zakharchenko, L. licea-Jimenez, S.A. Perez-Garcia, P. Vorobiev, U. Dehesa-carrasco, J.F. Perez-Robles, J. Gonzalez-Hernanadez, and Yu. Vorobiev, “Photovoltaic solar panel for a hybrid pv/thermal system”, *ELSEVIER Solar Energy material & Solar Cells*, vol. 82, pp. 253-261, 2004.

[4] B. Bekker, and H.J. Beukes, “Finding an optimal pv panel maximum power point tracking method”, presented at 2004 IEEE Africon. 7th Africon Conf. in Africa, Africa, Sept. 15-17, 2004.

[5] D. Yu, S. Yuvarajan, “Load Sharing in a hybrid power system with a pv panel and a pem fuel-cell”, presented at Ann. IEEE Conf. on Applied Power Electronics Conference and Exposition (APEC), Dallas, TX, USA March 1-23, 2006, doi: 10.1109/APEC.2006.1620698.

[6] D. Sera, R. Teodorescu, and P. Rodriguez, “PV panel model based on datasheet values”, presented at 2007 IEEE Int. Symposium on Industrial Electronics, Vigo, Spain, June 4-7, 2007, doi:10.1109/ISIE.2007.4374981.

[7] H. Kim, J. Kim, B. Min, D. Yoo, and H. Kim, “A highly efficient pv system using a series connection of dc-dc converter output with a photovoltaic panel”, *Renewable Energy*, vol. 34, pp. 2432-2436, 2009.

[8] N. Onat, and S. Ersöz “Fotovoltaik sistemlerde maksimum güç noktası izleyici algoritmalarının karşılaştırılması”, presented at V. Yenilenebilir Enerji Kaynakları Sempozyumu 2009, Diyarbakır, Türkiye, 2009.

[9] C. Liu, K.T. Chau, C. Diao, J. Zhong, X. Zhang, S. Gao , and D. Wu, “A new DC micro-grid system using renewable energy and electric vehicles for smart energy delivery,” presented at the 2010 IEEE Vehicle Power and Propulsion Conference, Lille, France, September 1-3, 2010.

[10] A.C. Brent, and D.E. Rogers, “Renewable rural electrification: Sustainability assessment of mini-hybrid off-grid technological systems in the African context,” *Renewable Energy*, vol. 35, no. 1, pp. 257-265, January 2010, doi: 10.1016/j.renene.2009.03.028.

[11] D. Yamegueu, Y. Azoumah, X. Py, and N. Zongo, “Experimental study of electricity generation by solar PV/diesel hybrid systems without battery storage for off-grid areas,” *Renewable Energy*, vol. 36, no. 6, pp. 1780-1787, June 2011, doi: 10.1016/j.renene.2010.11.011.

[12] W. Saranrom, and S. Polmai, “The efficiency improvement of series connected PV panels operating under partial shading condition by using per-panel DC/DC converter,” presented at the 8th Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand – Conference 2011, Khon Kaen, Thailand, May 17-19, 2011.

[13] N.C. Sahoo, I. Elamvazuthi, N.M. Nor, P. Sebastian, and B.P. Lim, “PV panel modelling using simscape,” presented at the 2011 International conference on Energy, Automation and Signal, Bhubaneswar, India, December 28-30, 2011.

[14] O. Palizban, M.A. Rezaei, and S. Mekhilef, “Active and reactive power control for a hybrid system with photovoltaic panel, wind turbine, fuel cells, electrolyzer and super capacitor in off-grid mode,” presented at 2011 IEEE International Conference on Control System, Computing and Engineering, Penang, Malaysia, November 25-27, 2011.

[15] Q. Mo, M. Chen, Z. Zhang, Y. Zhang, and Z. Qian, “Digitally controlled active clamp interleaved flyback converters for improving efficiency in photovoltaic grid-connected micro-inverter,” presented at the 2012 Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC), Orlando, FL, USA, February 5-9, 2012.

[16] A. Mehdipour, and H. Majdinasab, “Voltage –fed trans z source inverter in PV solar panel,” presented at 2012 5th International Conference on Computers and Devices for Communication (CODEC), Kolkata, India, December 17-19, 2012.

[17] A. Korodi, “Building a knowledge base to obtain the maximum power point for a PV panel,” presented at 2012 IEEE International Conference on Control Applications (CCA) Part of 2012 IEEE Multi-Conference on Systems and Control, Dubrovnik, Croatia, October 3-5, 2012.

[18] M. Kasper, D. Bortis, T. Friedli, and J. W. Kolar, “Classification and comparative evaluation of PV panel integrated DC/DC converter concepts,” presented at 15th International Power Electronics and Motion Control Conference (EPE-PEMC), Novi Sad, Serbia, September 4-6, 2012.

[19] H. Ghoddami, M.B. Delghavi, and A. Yazdani “An integrated wind-photovoltaic-battery system with reduced power-electronic interface and fast control for grid-tied and off-grid applications,” *Renewable Energy*, vol. 45, pp. 128-137, September 2012, doi: 10.1016/j.renene.2012.02.016.

[20] N. Altanneh, “Güneş pili ve hidrojen yakıt pilinden beslenen küçük bir elektrikli araç için batarya şarj sistemi tasarımı ve gerçekleştirilmesi,” M.S. thesis, Electrical and Electronics Engineering, Gazi University Ankara, Türkiye, 2012. [Online]. Available: https://tez.yok.gov.tr/UlusalTezMerkezi/

[21] R. Kale, S. Thale, and V. Agarwal, “Design and implementation of a solar PV panel integrated inverter with multi-mode operation capability,” presented at the 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC), Tampa, FL, USA, June 16-21, 2013.

[22] M. Edouard, and D. Njomo, “Mathematical modelling and digital simulation of PV solar panel using MATLAB software,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 9, pp. 24, September, 2013.

[23] S.A. Rahman, R.K. Varma, and T. Vanderheide, “Generalised model of a photovoltaic panel,” *IET Renewable Power Generation*, vol. 8, no. 3, pp. 217-229, April, 2014, doi:10.1049/iet-rpg.2013.0094.

[24] M.B. Ammar, M. Chaabene, and Z. Chtourou, “Artificial neural network based control for PV/T panel to track optimum thermal and electrical power,” *Energy Conversion and Management*, vol. 65, pp. 372-380, January, 2013. (ammar2013)

[25] C.-L. Shen, Y.-C. Lee, J.-C. Su, and C.-T. Tsai, “A high step-up DC/DC converter for PV panel application,” presented at the 2014 International Conference on Information Science, Electronics and Electrical Engineering, Sapporo, Japan, April 26-28, 2014.

[26] A.A.A. Hafez, “Multi-level cascaded DC/DC converters for PV applications,” *Alexandria Engineering Journal*, vol. 54, no. 4, pp. 1135-1146, December 2015.

[27] A. Belkaid, I. Colak, and K. Kayisli, “Implementation of a modified P&O-MPPT algorithm adapted for varying solar radiation condition, ” *Elektrical Engineering*, vol. 99, pp. 839-846, October 2016.

[28] A. Belkaid, I. Colak, K. Kayisli, “Optimum control strategy based on an equivalent sliding mode for solar systems with battery storage,” presented at the 2016 IEEE International Power Electronics and Motion Control Conference (PEMC), Varna, Bulgaria, September 25-28, 2016.

[29] Y.A. Alhammad, W.F. Al-Azzawi, and T.A. Tutunji, “Current control to improve COP of thermoelectric generatot and cooler for PV panel cooling,” presented at the 2016 13th International Multi-Conference on Systems, Signal & Devices (SSD), Leipzig, Germany, March 21-24, 2016.

[30] A. Belkaid, I. Colak, K. Kayisli, R. Bayındır, and H.I. Bulbul, “Maximum power extraction from a photovoltaic panel and a thermoelectric generator constituting a hybrid electrical generation system,” presented at the 2018 International Conference on Smart Grid (icSmartGrid), Nagasaki, Japan, December 4-6, 2018.

[31] S.R. Pendem, and S. Mikkili, “Modeling, simulation and performance analysis of PV array configurations (series, series- parallel and honey-comb) to extract maximum power under partial shading conditions, ” *Energy Reports,* vol. 4, pp. 274-287, November 2018.

[32] A. Belkaid, I. Colak, K. Kayisli, M. Sara, and R. Bayindir, “Modeling and simulation of polycrystalline silicon photovoltaic cells,” presented at the 2019 7th International Conference on Smart Grid (icSmartGrid), Newcastle, NSW, Australia, December 9-11, 2019.

[33] M.J. Mnati, V.G.M. Araujo, J.K. Abed, and A.V. Boosche, “Review different types of MPPT techniques for photovoltaic systems,” presented at the International Conference on Sustainable Energy and Environment Sensing (SEES 2018), Cambridge, United Kingdom, June, 2018.

[34] A. Belkaid, I. Colak, K. Kayisli, and R. Bayındır, “Improving PV system performance using high efficiency fuzzy logic control,” presented at the 2020 8th International Conference on Smart Grid (icSmartGrid), Paris, France, June 17-19, 2020.

[35] S. Beyarslan, “Yenilenebilir enerji kaynakları ile mikro şebeke tasarımı ve optimum çözümünün HOMER ile incelenmesi,” M.S. Thesis, Electrical Engineering, İstanbul Technical University, İstanbul, Türkiye, 2012. [Online]. Available: https://tez.yok.gov.tr/UlusalTezMerkezi/

[36] R. Boylestad, and L. Nashelsky, “Semiconductor diodes,” in *Electronic Devices and Circuit Theory,* 7th edition Upper Saddle River NJ, Columbus, Ohio, USA: Prentice Hall, 1998, ch. 1, pp. 1-50.

[37] N. Mohan, T.M. Undeland, and W.P. Robbins, “dc-dc switching- mode converters” in *Power Electronics Converters, applications, and design*, 2nd ed., USA, John Wiley & Sons, INC. 1995, ch. 7, pp. 161-199.

[38] J.D. Irwin, R. M. Nelms, and A. Patnaik, “Capacitance and Inductance,” in *Engineering Circuit Analysis,* 11th ed. Asia: John Wiley & Sons, 2015, ch. 6, sec 6.1 , pp. 232-238.

[39] Ahmet Gündoğdu, “System Identification based ARV-MPPT Technique for PV Systems Under Variable Atmospheric Conditions”, (2022), IEEE Access, Vol. 10, pp. 51325-51342.

[40] Reşat Çelikel, Musa Yılmaz, Ahmet Gündoğdu, “A voltage scanning-based MPPT method for PV power systems under complex partial shading conditions”, (2022), Renewable Energy, Vol. 184, pp. 361-373.

[41] Reşat Çelikel, Ahmet Gündoğdu, “Comparison of PO and INC MPPT Methods Using FPGA In-The-Loop Under Different Radiation Conditions”, (2021), Balkan Journal of Electrical and Computer Engineering, Vol. 9, No. 2, pp. 114-122 .

[42] Reşat Çelikel, Ahmet Gündoğdu, “System Identification-based MPPT Algorithm for PV Systems Under Variable Atmosphere Conditions Using Current Sensorless Approach”, (2020), International Transactions on Electrical Energy Systems, Vol. 30, No. 8, pp.1-21.

[43] A. Belkaid, J. Gaubert, A. Gherbi and L. Rahmani, "Maximum Power Point tracking for photovoltaic systems with boost converter sliding mode control," 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE), 2014, pp. 556-561, doi: 10.1109/ISIE.2014.6864673.

BIOGRAPHIES

Korhan Kayisli received a BSc degree in electronics education from Sakarya University, Sakarya, Turkey, in 2001, and an MSc degree in Electronics and Computer Science from Firat University, Elazig, Turkey, in 2004. He received PhD degree at the area of power electronics in Electric and Electronics Engineering at Firat University, Elazig, Turkey, in 2012. He worked as a research assistant between 2002 and 2012. He has worked in Firat University, Bitlis Eren University, Gelisim University, Nisantasi University, respectively. He is currently an assistant professor in the Department of Electrical Electronics Engineering, Engineering Faculty, Gazi University, Ankara, Turkey. He is an IEEE member and the co-editor of International Journal of Renewable Energy Research and International Journal of Engineering Science and Application. He also served as reviewer to many high ranked scientific journals. His fields of interest are power electronics, converter circuits, power factor correction, robust control, and educational technologies. He has published journal and conference papers on these areas. Additionally, he has worked as researcher in two EU mobility projects and other some projects.

Ruhi Zafer Caglayan received his Bachelor's degree in Electrical-Electronics Engineering from Gazi University, Ankara, Turkey in 2021 with English education. He is doing his Master of Science in Electrical and Electronics English at Gazi University, which he started in 2022.

Air Based Flexible Ultra-Thin Transparent ITO Based Broadband and Polarization Insensitivity Metamaterial Absorber

Gokhan Ozturk, Fatih Tutar, Mehmet Ertugrul, Abdulsemih Kockeser, and Yakup Ozturk

***Abstract*—In this study, a metamaterial-based transparent and flexible microwave absorber design was carried out. Transparent PET (polyethylene terephthalate) was used as the dielectric substrate and ITO (indium tin oxide) was used as the conductor for the air and metamaterial structure. The intended absorber provides %90 absorption in the range of 9.6 GHz to 34.8 GHz with a normal incidence angle of approximately 25.2 GHz. Oblique angle performance shows % 80 absorption up to 45 degrees. In addition, the designed absorber works as a polarization insensitive absorber as it provides the same absorption performance in both TE and TM polarization under the normal incidance of the electromagnetic wave. The transparent dielectric is only 2.85 mm thickness, making it thinner than comparable ultra-wideband transparent materials. The study was carried out as a simulation in the CST microwave simulator. The results obtained were compared with other reference studies.**

***Index Terms*—Transparent, ultra-thin, metamaterial absorber, flexible, wide-band**

I. Introduction

In recent years, metamaterials have increased their importance in microwave device design with their superior properties as a negative refractive index (n), high resonance and periodically manufacturing. Metamaterials are called left-handed materials because they propagate waves in the opposite direction of conventional materials. The first theoretical work on left-handed materials was carried out by Veselego. Veselego proved that mathematically the refractive index occurs simultaneously with the negative constitutive parameters [1]. Experimentally, Pendry fabricated the negative permittivity material with an electric field applied to infinitely thin wires, similar to the behavior of gases at plasma frequency. He also fabricated negative permeability with SRR (Spling Ring Resenator) [2], [3]. Later, these two structures were combined in one material and left-handed material was produced [4]. Besause of the unique feature of metamaterials, they use a lot of area such as optical lens [5], sensor [8], antenna [6], solar systems [7], invisibility cloak [9], polarization converters [11] and absorbers [10].

There are numerous absorber applications to use in civil and military radars. In these applications, it is aimed to reduce the radar cross sectional (RSC) in order to prevent the detection of the target. The first metamaterial absorber designed by Landy inspired the use of metamaterials as absorber[10]. Afterwards, metamaterial based absorbers have been diversified by considering features such as multi [12] or single layer [13], polarization insensivity [14], oblique angle performance [15] and thickness according to wavelength [16] in the literature. The absorbers designed using various geometries were made not only for the GHz frequency region [17] but also for the THz region [18]. In some metamaterial-based broadband absorber applications, SRR structure combined with lumped elements [19] and thin films [20] to overcome their bandwidth limitation. Metamaterial absorbers can also be classified in terms of the bandwidth they cover such as one [17], double [21], triple [22] or penta band [23]. In addition, in order to provide this wide band gap, wideband absorber designs have been carried out using transparent materials, since the selection of completely transparent materials will provide more advantageous situations like invisibility [24]. Various microwave absorbers have been designed using materials such as glass, PDMS, PVC and PET as optically transparent materials. Peng designed an absorber with %90 absorption in the 6.4-30 GHz band using water, PDMS and ITO transparent materials. He did not report the oblique angle performance of his absorber in his study [25]. Gao designed an absorber operating in the 14.4-30.4 GHz band using air, water and ITO transparent materials. The oblique angle performance was the same up to 30 degrees for both TE and TM polarizations, and the thickness of the absorber was only 0.184 λ [24]. In addition, although there is high oblique incidance performance for wider bandwidth in other transparent absorber studies, the thickness of the absorber was a disadvantage for thin applications [26], [27], [28]. In addition, the flexibility of the absorbers was reported as a valnerable feature in studies. In this study, we proposed completely transparent ITO-based metamaterial

Gokhan OZTURK is with the Department of Electrical and Electronics Engineering, Ataturk University, 25100 Erzurum, Turkey (e-mail: gokhan.ozturk@atauni.edu.tr),

https://orcid.org/0000-0001-8106-0053

Fatih TUTAR is with the Department of Electrical and Electronics Engineering, Ataturk University, 25100 Erzurum, Turkey
(e-mail: fatihtutar00@gmail.com),

https://orcid.org/000-0003-0668-3319

Mehmet ERTUGRUL is with the Department of Electrical and Electronics Engineering, Ataturk University, 25100 Erzurum, Turkey (e-mail: ertugrul@atauni.edu.tr)

https://orcid.org/0000-0003-1921-7704

Abdulsemih KOÇKESER is with the Department of Electrical and Electronics Engineering, Ataturk University, 25100 Erzurum, Turkey (e-mail: abdulsemihkockeser@gmail.com)

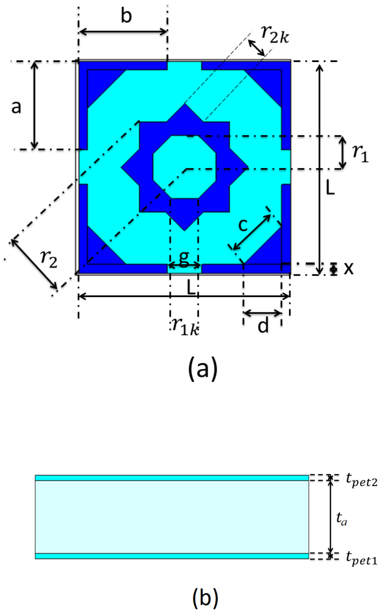
https://orcid.org/0000-0002-8222-3536

Yakup OZTURK is with the Department of Ministry of Education, 25100 Erzurum, Turkey (e-mail: yakupozturk25@hotmail.com)

https://orcid.org/0000-0001-6830-5837

Manuscript received August 23, 2021; accepted July 20, 2022.

DOI: 10.17694/bajece.986271



(a)

(b)

Fig. 1. (a) Front profile and (b) side profile of proposed metamaterial based transparent absorbent design

absorber design which operate between the 9.6-34.8 with 25.2 GHz bandwidth approximately and cover the X-, Ku-, K- and Ka- bands with 90% absorption performance. The design consists of air, PET and ITO, which are flexible and transparent. Absorber acts as polarization insensivity, because the proposed design shows the same absorption performance in both TE and TM polarization. The thickness of the intended absorber is only $2.85mm(0.21\lambda)$ and it show ultra-thin feature compared to other transparent metamaterial absorbers in the literature. In Section II, the design of the absorber and the theoretical background are given to understand the absorption mechanism. Absorption, impedance and constitutive parameters were obtained by using the simulation results in Section III. Finaly, the oblique incidance performance, relative bandwidth (RB), flexiblity, thickness of absorber were compared with other reference studies in Section IV.

II. Theory and Design

The proposed transparent metamaterial absorber is as shown in Fig. 1 (a-b). As seen in Fig. 1 (a-b), the bottom and top parts of the three-layer structure covered with ITO material, which is a transparent and conductive thin film. The battom level consists entirely of ITO with a thickness of $100nm$ and $10\Omega/m^2$ resistance. As seen in Fig. 1 (a), the top part consist of metamaterial coated by ITO with $100nm$ thickness and $35\Omega/m2$ resistance (the parts shown with dark blue). Dimensions of the ITO structure shown in Fig. 1 (a); $L = 12.2$ mm, $r_1 = 1.83$ mm, $r_2 = 3.2$ mm, $a = 5.10$ mm, $b = 5.10$ mm, $x = 0.50$ mm, $c = 3.14$ mm, $d = 4.6$ mm, $g = 2$ mm, $r_{1k} = 1.52mm$ and $r_{2k} = 1.52mm$, respectively.

An part of the lowest ITO coated level and a lower level of the uppermost ITO coated metamaterial structure consists of PET as Fig. 1 (b). The dielectric constant of PET is $\varepsilon_r = 3.2$ and the loss tangent is $\sigma = 0.003$. The middle part consists of an air medium with a dielectric $\varepsilon_r = 1$ and a thickness of $d = 2.5$ mm. The thickness of the upper level and the lowest level PET has a thickness of $t_{pet1} = t_{pet2} = 0.175mm$ and air thickness is $t_a = 2.5mm$. The metamaterial structure at the top consists of ITO with a conductivity of $35\Omega/m2$ and a thickness of $100nm$. The design given in Fig. 1 consists of transparent and flexible, which are air, water and PET. Therefore, the design can provide optically transparenty and flexibility advantage. Depending on the scattering parameters, the microwave absorption performance can be obtained as follows [27]

$$Absorbtion = 1 - R(\omega) - T(\omega), \tag{1}$$

Here $R(\omega) = |S_{11}|^2$ and $T(\omega) = |S_{21}|^2$, where S_{11} and S_{21} are reflection and transmission scattering parameters, respectively. Impedance matching is an important criterion to achieve good performance of the absorber. Using the scattering parameters, the normalized wave impedance of the absorber can be obtained as follows [30]

$$\bar{z} = \sqrt{\frac{\mu_{eff}}{\epsilon_{eff}}} = \sqrt{\frac{(1+S_{11}^2) - S_{21}^2}{(1-S_{11}^2) - S_{21}^2}}, \tag{2}$$

Here $\bar{z}$, ε_{eff} and μ_{eff} are normalized wave impedance, electrical permittivity and magnetic permeability, respectively. The electric and magnetic dielectric constants of the absorber directly affect the normalized impedance as seen in equation 2. The permittivity and permeability of the metamaterial absorber can give information about how the absorber works and performance of absorber. Depending on the scattering parameters, the permittivity and permeability of the absorber can be obtained as follows [30]

$$\epsilon_{eff} = 1 + \frac{2jS_{11} - 1}{k_o d S_{11} + 1}, \tag{3}$$

$$\mu_{eff} = 1 + \frac{2jS_{11} + 1}{k_o d S_{11} - 1}, \tag{4}$$

Where, k_o is the wave number of the free-space and d is the thickness of the absorber. Since the bottom of the proposed absorber is covered with high conductivity ITO material, ε_r and μ_r can be roughly obtained by assuming that the S_{21} parameter goes to about 0 in equations 3 and 4. For another analysis of the metamaterial absorber physical infrastructure and working mechanism, the equivalent circuit based on transmission line theory is given in Fig.1. Impedance Z_a refers to the part below the metamaterial in Fig. 2 and is expressed as

$$Z_a = j\frac{Z_0}{\sqrt{\epsilon_r}} \tan\frac{2\pi f\sqrt{\epsilon_r}d}{c}, \tag{5}$$

where j is the imaginary part, d is the thickness of the substrate and ϵ_r is the relative permittivity, f the frequency of the incident wave, c is the velocity of light. Z_b, which is the impedance of the metasurface, can be written as [29]

$$Z_b = R + j(2\pi f L - \frac{1}{2\pi C}), \tag{6}$$

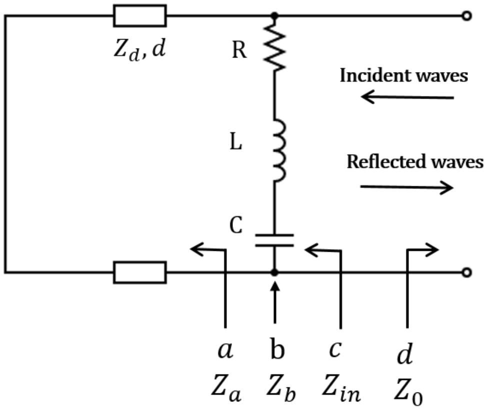


Fig. 2. RLC Equivalent circuit of proposed metamaterial based transparent absorber

The total input impedance of port c, which includes impedances of Z_a and Z_b, is given with Z_{in} as follow

$$Z_{in} = \frac{Z_a Z_b}{Z_a + Z_b}, \quad (7)$$

Port d represents the characteristic impedance of free space and is denoted by Z_o and its value is 377Ω .

$T(\omega)$ mentioned in equation 1 is the transmission scattering coefficient and it can be taken as approximately zero since the back of the design below shows metal feature and $R(\omega)$ expressed as follow

$$R(\omega) = |\frac{Z_{in} - Z_0}{Z_{in} + Z_0}|^2, \quad (8)$$

The matching of free space and total impedance Z_{in} has to be equal Z_0. Equations 9 and 10 are obtained by using equations 5-8.

$$\frac{R}{R + (2\pi f L - \frac{1}{2\pi C})} = \frac{1}{Z_0}, \quad (9)$$

$$\frac{(2\pi f L - \frac{1}{2\pi C})}{R^2 + (2\pi f L - \frac{1}{2\pi C})^2} = \frac{\sqrt{\epsilon_r}}{Z_0} \cot(\frac{2\pi f \sqrt{\epsilon_r} d}{c}), \quad (10)$$

The mathematical expression of R, which is in the equivalent circuit by means of equations 9 and 10, is expressed as follow

$$R = \frac{Z_0 \tan^2(\frac{2\pi f \sqrt{\epsilon_r} d}{c})}{\epsilon_r + \tan^2(\frac{2\pi f \sqrt{\epsilon_r} d}{c})}, \quad (11)$$

Besides, relationship between equivalent parameters in the Fig. 2 as follow

$$(2\pi f L - \frac{1}{2\pi f C}) = -\frac{\sqrt{\epsilon_r} Z_0 \tan(\frac{2\pi f \sqrt{\epsilon_r} d}{c})}{\epsilon_r + \tan^2(\frac{2\pi f \sqrt{\epsilon_r} d}{c})}, \quad (12)$$

Where, L and C are equivalent inductance and equivalent capacitance, respectively.

III. Simulation Results

In order to determine S-parameters in the CST simulator, frequency domain is selected in software and unitcell boundary conditions are applied in flouquet mode for boundary condition. The hexahedral mesh type was chosen for precision analysis. 20x20 mesh cells were set to the unitcell metasurface. Accuracy set to $1e^{-12}$ and solver order high accuracy in frequency domain solver. After applying the boundary conditions in CST, the S-parameters results were obtained for TE and TM mode as given in Fig. 3. As seen in Fig. 3, reflected waves in the range of 9.6 GHz to 34.8 GHz for both modes are below $-10dB$. The $x-y$ plane symmetrical geometry of the proposed design ensures that the same S-parameters are obtained for both TE and TM modes as polarization insensitivity. Also, at the back of the design, there is metal termination by choosing a very small resistive ITO surface $(10\Omega/m^2)$. This surface acts just like metal termination with good conductor. Thus, the transmission scattering parameter (S_{21}) approached almost zero, providing transmission at approximately $-20dB$. In Fig. 3, design shows strong resonance and absorbation at 11.73 GHz, 25.84 GHz, 29.71 GHz and 32.85 GHz. The S11 parameter drops sharply at 30.56 GHz, where the reflection is minimum. Fig. 4 show absorbtion results under normal incidence for both TE and TM polarizations. The absorption results are the same in both polarization due to the symmetrical design. As seen in Fig. 4, more than % 90 absorption was achieved from 9.6 to 34.8 GHz. Maximum absorption with % 99.99 were achieved at the 29.71 GHz frequency, because there is a minimum reflection in this frequency. By using the S-parameters to equation 1, oblique incidence absorption performance for TE and TM polarization are presented with % 80 absorption in Fig. 5 (a-b). As it can be seen in Fig. 5 (a-b), if incidence angle increase, absorption performance decreases, gradually. Absorption performance of the design under oblique incidence is up to 45^o. Especially, angle sensitivity of absorber for TM mode is more than TE mode between $30^o - 45^o$. For strong impedance matching, the normalized impedance of the proposed absorber is expected to approach the normalized impedance of air. The real and imaginary part of the normalized impedance obtained using equation 2, the real and imaginary parts of the electric and magnetic permeability using equation 3-4 are given in Fig. 6 (a-b-c). As can be seen from the impedance of the absorber in Fig. 6 (a), between 9.6-34.8 GHz, the z_{real} expression oscillates around 1 value, while the z_{imag} expression oscillates around 0. Fig. 7 shows the surface current distributions for four resonance

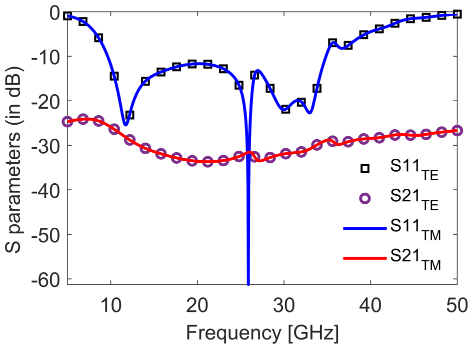


Fig. 3. S-parameters of the proposed metamaterial-based transparent absorber design in dB for TE and TM mode under normal incidence.

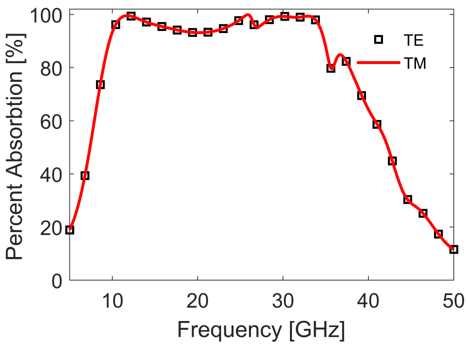


Fig. 4. Percent absorption performance of the proposed metamaterial-based transparent absorbent design for TE and TM mode under normal incidence.

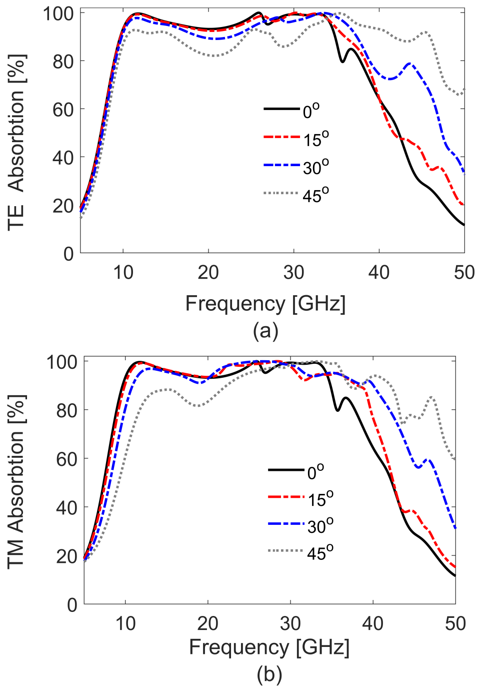


Fig. 5. (a) Percent absorption performance under oblique incidance for TE mod and (b) Percent absorption performance under oblique incidance for TM mod of the proposed metamaterial-based transparent absorber design.

points which are 11.73 GHz, 25.84 GHz, 29.71 GHz and 32.85 GHz. The 11.73 GHz is the first resonance frequency and top and bottom surface currents distribution of 11.73 GHz resonance frequency are shown in Fig. 7 (a). When looking at the directions of the surface current distributions in the top and bottom parts, it is seen that they are in opposite directions, so magnetic resonance has occurred at this resonance point. The top and bottom surface currents distributions of the 25.84 GHz resonance frequency are shown in Fig. 7 (b), when looking at the directions of the surface current distributions in the top and bottom parts, it is seen that they are in the same direction, so electrical resonance has occurred at this resonance point. Likewise, the top and bottom surface currents of 29.71 and 32.85 GHz resonance frequencies are shown in Fig. 7 (c) and Fig. 7 (d), when looking at the directions of the surface current distributions in the top and bottom parts, it is seen that they are in the same direction, so electrical resonance has occurred at these two resonance points.

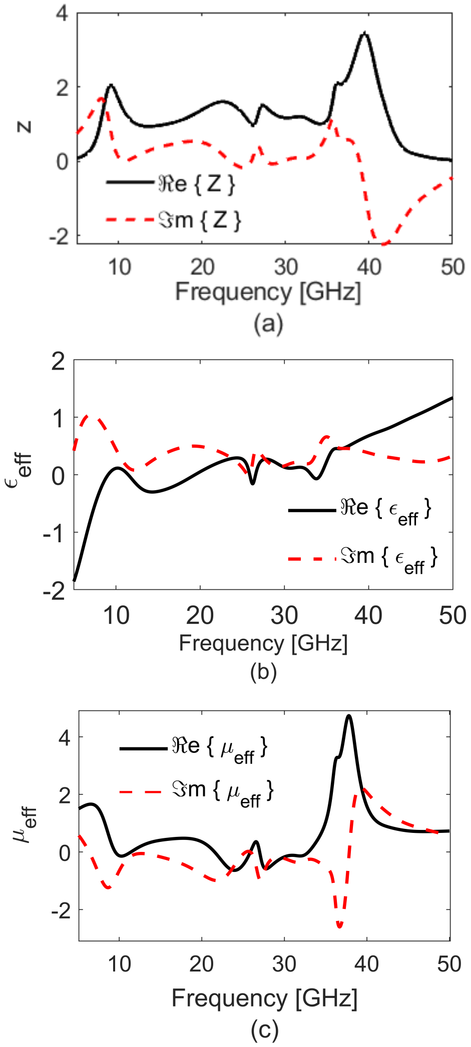


Fig. 6. (a) Real and imaginary parts of normalized wave impedance (b) real and imaginary parts of effective permittivity and (c) Real and imaginary parts of effective permeability of the proposed metamaterial-based transparent absorber design.

We analyzed the absorption performance with the loss tangent variation of the proposed absorber in the Fig. 8. As seen in the figure, the absorption performance of the proposed design decreases when the loss tangent increases too much. We also analyzed the absorption performance by varying the dielectric thickness of the proposed absorber for $t_a = 1.0$ mm, $t_a = 2.5$ mm and $t_a = 3.5$ mm. As seen in the Fig. 8, the performance of the absorber is optimum for $t_a = 2.5$ mm.

Table I give information about the performance of our and other optical transparent metamaterial absorber in the literature. The performance analysis of the study is presented based on some parameters such as the bandwidth, the relative bandwidth (RB), material thickness, flexibility in Table I. As can be seen in Table I, the proposed transparent metamaterial-based absorber is more useful for wide band applications with

TABLE I
PERFORMANCE OF THE PROPOSED ABSORBER RELATIVE TO OTHER REFERENCED STUDIES.

Ref.	Op. BW [GHz]	Flex	Angle	Thickness	RB
Ref [25]	6.4-30 GHz (23.6 GHz)	No	-	$4.5mm(0.273\lambda)$	$\%129.67$
Ref [24]	14.4-30.4 GHz (16 GHz)	No	30^o	$3.176mm(0.184\lambda)$	$\%72.07$
Ref [26]	14.4-33.7 GHz (19.3 GHz)	No	40^o	$6.4mm(0.513\lambda)$	$\%101.31$
Ref [27]	8-18 GHz (10 GHz)	Yes	45^o	$4.5mm(0.22\lambda)$	$\%76.92$
Ref [28]	5.61-29.17 GHz (23.5 GHz)	Yes	60^o	$4.5mm(0.273\lambda)$	$\%135.7$
Prop. Absorb.	9.6-34.8 GHz (25.2 GHz)	Yes	45^o	$2.85mm(0.21\lambda)$	$\%113.5$

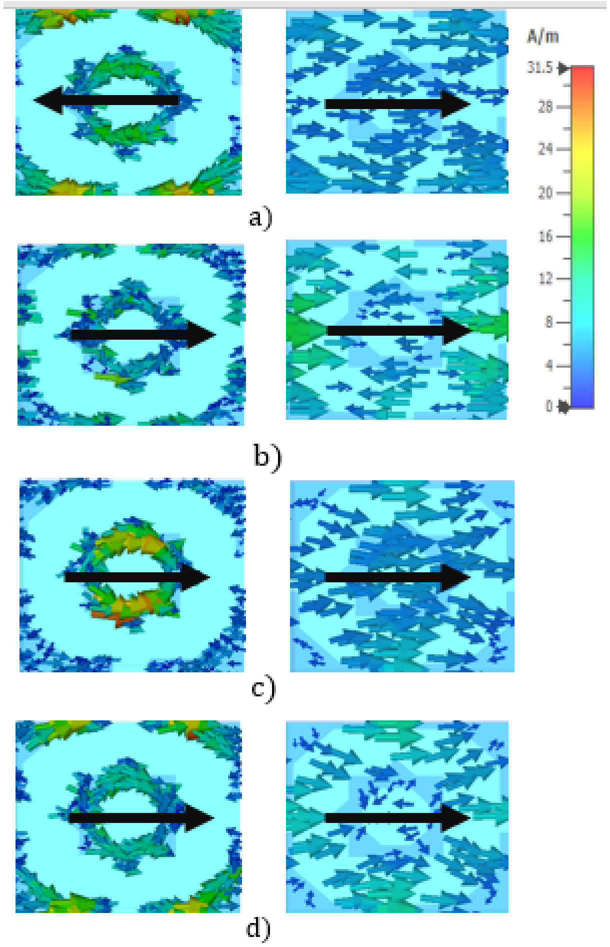


Fig. 7. Surface current distributions at the top and ground layers corresponding to a) 11.73 GHz b) 25.84 GHz c) 29.71 GHz d) 32.85 GHz

25.2 GHz bandwidth than references [24], [25], [26], [27], [28] and more competitive for ultra thin applications with $2.85mm(0.21\lambda)$ thickness in comparison with [24], [25], [26], [27], [28]. Besides, although referances [24], [25], [26] have a non-flexibility disadvantage, our proposed design show the flexibility performance because of selecting the flex materials. Since our design based on the air, we achieved good performance less material compared to [24], [25], [26], [27], [28]. The oblique incidence performance of absorber is up to (45^o) with wide angle relatively to other studies [24], [25], [26]. The RB performance with $(\%113.5)$ is more than references [24], [26], [27].

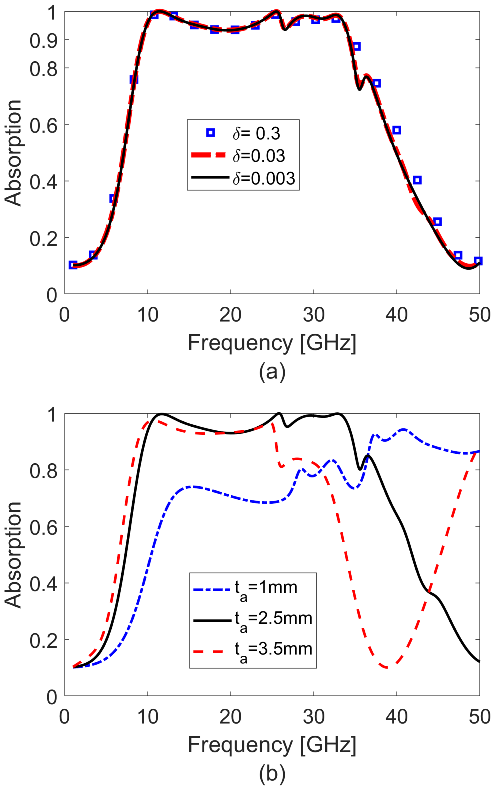


Fig. 8. The performance of the proposed absorber (a) according to the loss tangent variation (b) according to the thickness of the substrate

IV. CONCLUSION

In this study, a metamaterial-based transparent, flexible and ultra-thin absorber design is proposed. For the metamaterial structure, transparent thin film (ITO) and flexible transparent materials (air, PET). The proposed absorber design works for a very wide bandwidth (25.2 GHz) between 9.6 GHz and 34.8 GHz with $\%90$ absorption. The performance of the design is the same for both TE and TM polarized incoming waves, and the structure works as a polarization insensitive microwave absorber. In addition, the absorber shows an absorption performance of more than $\%$ 80 up to 45^o in both polarization at oblique incidence. Compared to other transparent metamaterial-based ultra wide band absorbers in the literature, it exhibits more comfortable properties in terms of bandwidth, thickness, angle and flexibility.

V. ACKNOWLEDGMENT

This work is partially supported by the Scientific and Technological Research Council of TURKEY (TUBITAK) under the Project Number 218M341, Fatih TUTAR was supported by The Scientific and Technological Research Council of Turkey (TÜBITAK), Directorate of Science Fellowships and Grant Programmes (BİDEB) and The Higher Education Council of Turkey (YÖK) under a 100/2000 doctorate scholarship.

REFERENCES

[1] V. G. Veselego, "The Electrodynamics of Substances with Simultaneously Negative Values of ϵ and μ," *Physics-Uspekhi*, vol. 10, no. 4, pp. 509–514, 1968.

[2] J. B. Pendry, A. J. Holden, D. J. Robbins and W. J. Stewart, "Low frequency plasmons in thin-wire structures," *Journal of Physics: Condensed Matter*, vol. 10, no. 22, pp. 4785–4809, Mar. 1998.

[3] J. B. Pendry, A. J. Holden, D. J. Robbins and W. J. Stewart, "Magnetism from conductors and enhanced nonlinear phenomena," *Journal of Physics: Condensed Matter*, vol. 47, no. 11, pp. 2075–2084, Nov. 1999.

[4] D. R. Smith, W. J. Padilla, D. C. Vier, S. C. Nemat-Nasser and S. Schultz, "Composite medium with simultaneously negative permeability and permittivity," *Physical review letters*, vol. 84, no. 18, pp. 4184–4187, May. 2000.

[5] N. Kundtz and D. R. Smith, "Extreme-angle broadband metamaterial lens," *Nature Materials*, vol. 9, no. 2, pp. 129–132, Feb. 2010.

[6] R. W. Ziolkowski and A. Erentok, "Metamaterial-based efficient electrically small antennas," *IEEE Transactions on antennas and propagation*, vol. 54, no. 7, pp. 2113–2130, Jul. 2006.

[7] H. Wang, V. P. Sivan, A. Mitchell, G. Rosengarten, P. Phelan and L. Wang, "Highly efficient selective metamaterial absorber for high-temperature solar thermal energy harvesting" *Solar Energy Materials and Solar Cells*, vol. 137, pp. 235–242, Feb. 2015.

[8] W. Withayachumnankul, K. Jaruwongrungsee, A. Tuantranont, C. Fumeaux and D. Abbott, "Metamaterial-based microfluidic sensor for dielectric characterization," *Sensors and Actuators A: Physical*, vol. 189, no. 20, pp. 233–237, Jan. 2013.

[9] D. Schurig, J. J. Mock, B. J. Justice, S. A. Cummer, J. B. Pendry, A. F. Starr and D. R. Smith, "Metamaterial electromagnetic cloak at microwave frequencies," *Science*, vol. 314, no. 5801, pp. 977–980, Oct. 2006.

[10] N. I. Landy, S. Sajuyigbe, J. J. Mock, D. R. Smith and W. J. Padilla, "Perfect metamaterial absorber," *Phys. Rev. Lett.*, vol. 100, no. 20, pp. 207402, May. 2008.

[11] A.K. Fahad, C. Ruan, S.A. Ali, R. Nazir, T.U. Haq, S. Ullah and W. He, "Triple-wide-band Ultra-thin Metasheet for transmission polarization conversion," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, Jun. 2020.

[12] D.E. Wen, H. Yang, Q. Ye, M. Li, L. Guo and J. Zhang, "Broadband metamaterial absorber based on a multi-layer structure," *Physica Scripta*, vol. 88, no. 1, pp. 015402 Jul. 2013.

[13] P. Nochian and Z. Atlasbaf, "A Novel Single Layer Ultra-Wideband Metamaterial Absorber," *Progress In Electromagnetics Research Letters*, vol. 93, pp. 107–114 Jul. 2020.

[14] M.Q. Dinh, T. Le Hoang, H.T. Vu, N.T. Tung and M.T. Le, "Design, fabrication, and characterization of an electromagnetic harvester using polarization-insensitive metamaterial absorbers," *Journal of Physics D: Applied Physics*, vol. 54, no. 34, pp. 345502 Aug. 2021.

[15] J. Wang, S. Qu, Z. Xu, H. Ma, Y. Yang, C. Gu and X. Wu, " A polarization-dependent wide-angle three-dimensional metamaterial absorber," *Journal of magnetism and magnetic materials*, vol. 321, no. 18, pp. 2805–2809 Apr. 2017.

[16] Y. Zhu, K. Donda, S. Fan, L. Cao and B. Assouar, " Broadband ultra-thin acoustic metasurface absorber with coiled structure," *Applied Physics Express*, vol. 12, no. 11, pp. 114002 Nov. 2019.

[17] G. Sen, A. Banerjee, A. Nurul Islam and S. Das, " Ultra-thin miniaturized metamaterial perfect absorber for x-band application," *Microwave and Optical Technology Letters*, vol. 58, no. 10, pp. 2367–2370 Oct. 2016.

[18] J.W. Park, D.L. Vu, H.Y. Zheng, J.Y. Rhee, K.W. Kim and Y.P. Lee, " THz-metamaterial absorbers," *Advances in Natural Sciences: Nanoscience and Nanotechnology*, vol. 4, no. 1, pp. 015001 Mar. 2013.

[19] T.K.T. Nguyen, T.N. Cao, N.H. Nguyen, D.T. Lee, X.K. Bui, C.L. Truong and T.Q.H. Nguyen, "Simple design of a wideband and wide-angle insensitive metamaterial absorber using lumped resistors for X-and Ku-bands," *IEEE Photonics Journal*, vol. 13, no. 3, pp. 015001 Jun. 2021.

[20] R. Zvagelsky, D. Chubich, A. Pisarenko, Z. Bedran and E. Zhukova, "Plasmonic Metasurfaces as Surface-Enhanced Infrared Absorption Substrates for Optoelectronics: Alq3 Thin-Film Study," *The Journal of Physical Chemistry C*, vol. 125, no. 8, pp. 4694–4703 Mar. 2021.

[21] Y. Zhang, J. Lv, L. Que, Y. Zhou, W. Meng and Y. Jiang, "A double-band tunable perfect terahertz metamaterial absorber based on Dirac semimetals," *Results in Physics*, vol. 15, no. 102773, Dec. 2019.

[22] J. Wang, X. Wan and Y. Jiang, "Tunable Triple-Band Terahertz Absorber Based on Bulk-Dirac-Semimetal Metasurface," *IEEE Photonics Journal*, vol. 13, no. 4, pp. 1–5, Aug. 2021.

[23] V.B. Shalini, " A polarization insensitive miniaturized pentaband metamaterial THz absorber for material sensing applications," *Optical and Quantum Electronics*, vol. 53, no. 5, pp. 1–14, May. 2021.

[24] Z. Gao, Q. Fan, X. Tian, C. Xu, Z. Meng, S. Huang and C. Tian, "An optically transparent broadband metamaterial absorber for radar-infrared bi-stealth," *Optical Materials*, vol. 112, no. 110793, Feb. 2021.

[25] Y. Pang, Y. Shen, Y. Li, J. Wang, Z. Xu and S. Xu, "Water-based metamaterial absorbers for optical transparency and broadband microwave absorption," *Journal of applied physics*, vol. 123, no. 15, pp. 155106, Apr. 2018.

[26] F. Lu and T. Han, "Optically Transparent Ultra-broadband Metamaterial Absorber," *2019 Photonics and Electromagnetics Research Symposium-Fall (PIERS-Fall)*, pp. 2592–2595, Dec. 2019.

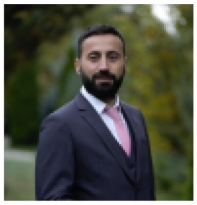
[27] Q. Zhou, X. Yin, F. Ye, R. Mo, Z. Tang, X. Fan and L. Zhang, "Optically transparent and flexible broadband microwave metamaterial absorber with sandwich structure," *Applied Physics A*, vol. 125, no. 2, pp. 131, Feb. 2019.

[28] Y. Zhou, S. Li, Y. Jiang, C. Gu, L. Liu and Z. Li, "An ultra-wideband and wide-angle optically transparent flexible microwave metamaterial absorber," *Journal of Physics D: Applied Physics*, vol. 54, no. 27, pp. 275101, Jul. 2021.

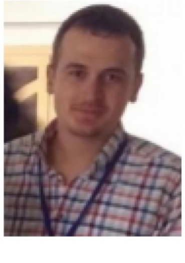
[29] R. Deng, M. Li, B. Muneer,Q. Zhu, Z. Shi, L. Song and T. Zhang, "Theoretical Analysis and Design of Ultrathin Broadband Optically Transparent Microwave Metamaterial Absorbers," *Materials*, vol. 11, no. 107, pp. 1-15, Jan. 2018.

[30] G. Ozturk, "Triple Band Wide Angle Polarization Insensitive Metamaterial Absorber," *Journal of Science and Technology*, impress, Aug. 2021.

BIOGRAPHIES



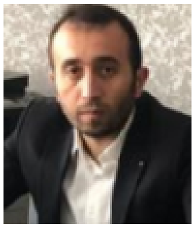
Gokhan OZTURK received the B.Sc. degree in electrical and electronics engineering from Fırat University, Elazıg, Turkey, in 2009, and the M.Sc. and Ph.D. degrees in electrical and electronics engineering from Ataturk University, Erzurum, Turkey, in 2014 and 2018, respectively. He was a Research Assistant with the Department of Electrical and Electronics Engineering, Igdır University, Igdır, Turkey, from 2010 to 2012, Kafkas University, Kars, Turkey, in 2012, and the Department of Electrical and Electronics Engineering, Ataturk University, from 2012 to 2018, respectively. Since 2018, he has been an Assistant Professor with the Department of Electrical and Electronics Engineering, Ataturk University. His current research interests include the characterization of material by microwaves, meta-materials, and numerical methods in electromagnetic.



Fatih TUTAR received the B.Sc. degree in electrical and electronics engineering from Ataturk University, Erzurum, Turkey, in 2018, and the M.Sc. degree in electrical and electronics engineering from Ataturk University, Erzurum, Turkey, in 2021. His current research interests include meta-materials, microwave polarization converters and absorbers in electromagnetic.

Mehmet ERTUGRUL received the B.Sc. degree from the Department of Physics, Ataturk University, Erzurum, Turkey, in 1986, and the M.Sc. and Ph.D. degrees in atomic physics from Ataturk University in 1990 and 1994, respectively. From 1994 to 1996, 1996 to 2001, and 2001 to 2002, he was an Assistant Professor, an Associate Professor, and a Full Professor with the Department of Physics, Ataturk University, respectively, where he has been a Full Professor with the Department of Electrical and Electronics Engineering since 2003. He is the author or a coauthor of more than 120 articles published in international journals. His current research interests include superconducting and semiconducting devices with applications, nanofabrication, and nanoelectronics. Prof. Ertugrul was a recipient of the award by The Scientific and Technological Research Council of Turkey and the Turkish Academy of Sciences.

Abdulsemih KOÇKESER received the B.Sc. degree in electrical and electronics engineering from Ataturk University, Erzurum, Turkey, in 2021, His current research interests include meta-materials and microwave absorbers.



Yakup OZTURK received the B.Sc. degree in faculty of education from Ataturk University, Erzurum, Turkey, in 2004, His current research interests include computer and informatics.

Effect of Different Kernel Functions on Hazardous Liquid Detection Using a New Spectroscopy System and Support Vector Machines

Ebru Efeoglu and Gurkan Tuna

Abstract—**Spectroscopy methods have become widespread in many applications including liquid classification. In this study, a new spectroscopy system that can classify liquids without opening the lid of their containers is proposed. Thus, the operators are prevented from being exposed to harmful substances and wasting time. Everyday liquids such as carbonated drinks, fruit juices, shampoo, cream and alcoholic beverages and hazardous liquids were characterized remotely by the method in which spectroscopic signatures of a total of 52 liquids were used. In order to be able to classify liquids with the highest accuracy, it is also important to determine the most suitable measurement system as well as the correct selection of the classification algorithm and algorithm parameters that show the best performance. In this study, Support Vector Machines algorithm, which is a very successful algorithm in separating binary classes, is used. In addition, the effects of the algorithm on the classification performance have been examined using different kernel functions and cross-validation technique has been used for the performance analysis. As a result of the performance analysis, it is seen that up to 100% success can be achieved when linear or polynomial kernel functions have been preferred.**

Index Terms—**Hazardous liquids, Support Vector Machines, Kernel functions, Accuracy.**

I. Introduction

SOME FLAMMABLE liquids that are readily available can be used to make explosives. Examples of these liquids are acetone and alcohol types. By mixing acidic drinks with acetone, a very powerful handmade explosive substance called TATP (Tricycloacetonperoxide) can be obtained. Bottles, which are part of our daily life, can be used to store such liquids. For this reason, methods for non-contact detection of liquids in any container are of great importance. Many methods have been developed to detect hazardous liquids at security checkpoints. The relationship between ethanol concentration in alcohol-water solutions at different concentrations and terahertz reflection signals was demonstrated using terahertz (THz) time domain spectrometer [1], and the detection of flammable and explosive liquids was realized [2][3]. The THz transmission spectra of explosives and recent developments in spectroscopic techniques for detection of explosives were investigated using a THz-TDS and Optical Parametric Oscillator based system in [4]. Similarly, Nuclear Magnetic Resonance (NMR) method was used to analyze unknown liquids together with infrared (IR) spectroscopy and classical chemical color tests in [5]. As given in [6].NMR method was also used to detect and classify liquid explosives. On the other hand, in [7] Raman spectroscopy method was proposed for non-contact detection of hazardous liquids stored in glass and plastic containers. Similarly, to classify liquids in glass bottles the use of a low energy X-ray transmission system was proposed in [8].

Microwave spectroscopy methods are more practical and less costly than other methods [9]. These features of the method have recently increased the interest in the method. Dielectric properties of liquids can be analyzed with the method of dielectric spectroscopy. Different measurement techniques have been developed for this purpose and several research studies have been carried out in this domain in recent years. Simulation studies were carried out to measure the permeability of liquids using an open-ended microwave waveguide [10]. Static dielectric permittivity ($\varepsilon0$) and relaxation time (τ) was acquired by the least-square-fit method in [11]. Dielectric relaxation mechanisms and the temperature dependence of complex permeability of water were calculated in [12]. The relationship between water molecules and dielectric was deduced in [13]. An optimized rectangular waveguide cavity resonator was designed for compositional analysis of liquid solutions in [14]. Cooking oil classification was made using dielectric spectroscopy at 8.2-12.1 GHz microwave frequencies [15]. Microwave spectroscopy method was used not only in liquid classification but also in determining the quality parameters of tomato paste [16] and silicone [17]. Coaxial probe method was the most commonly

EBRU EFEOGLU is with Kütahya Dumlupınar University, Kütahya, Turkey, (e-mail: ebru.efeoglu@dpu.edu.tr).

https://orcid.org/0000-0001-5444-6647

GURKAN TUNA is with Trakya University, Edirne, Turkey, (e-mail: gurkantuna@trakya.edu.tr).

https://orcid.org/0000-0002-6466-4696

Manuscript received July 27, 2021; accepted July 29, 2022.
DOI: 10.17694/bajece.975050

used microwave measurement method is the past and it was used for different purposes including the dielectric measurement of biological [18] in the diagnosis of breast cancer [19]. However, in recent years remote and non-contact measurement and machine learning methods have been gaining interest. Machine learning techniques were used in toxic liquid detection with a thick film gas sensor [20]. Performance analysis of a 2% Fe2O3-added thick film gas sensor was performed in toxic liquid detection using machine learning techniques [21].

In this study, an antenna was designed and it was connected to the Vector Network Analyzer (VNA) to collect spectroscopic signatures of liquids in the microwave frequency band. Then, Support Vector Machine (SVM) algorithm was applied to these spectra consisting of hazardous and everyday liquids to classify hazardous liquids. In addition, the effects of using different kernel functions on the success of the SVM algorithm were examined using various performance criteria.

II. EXPERIMENTAL SETUP AND METHODOLOGY

The complex dielectric permeability value varies depending on the chemical composition of the liquid tested in the proposed microwave spectroscopy method. This change in dielectric permeability affects the electromagnetic response of the antenna, and the different electromagnetic responses of liquids enable us to obtain information about the liquid. Measurement system used in this study consists of a VNA and a patch antenna, as shown in Fig.1. From the VNA, a signal is sent to the liquid in the 1.42-1.53 GHz frequency range of the microwave band and the amplitude of the signal reflected from the liquid is measured. The interaction between molecules and microwaves causes the molecules to rotate and align with the electromagnetic field. Polarization and depolarization of molecules in liquids with different dielectric permeability values, and energy loss due to friction of the directing molecules in the wave velocity cause a decrease in the magnitude of the wave. Dielectric loss factor measures the efficiency of energy loss [22]. The signal amplitude is a function of the dielectric constant of the sample and the change in dielectric loss factor [23]. The antenna was formed by placing a circular geometry conductive layer on a dielectric layer on a ground plane. This conductive layer provides the absorption or radiation of electromagnetic waves. Copper is used as the conductive layer. Rjadiation occurs between the conductive layer of the antenna and the ground plane and the most radiation occurs in the edge areas of the conductive layer. The reason why the circle patch was preferred is the symmetrical radiation characteristic of the circular patch which is not found in other types of patches. Another factor contributing to the circular patch selection was the use of a circular bottle for measurements. The thickness of the dielectric layer is directly proportional to the frequency bandwidth. The dielectric constant of the dielectric layer in the designed antenna is 4.4. Its dimensions are 10x10 cm and its thickness is 1.6 mm. The coaxial probe method was used as the feeding method for the designed antenna because the coaxial probe feeding method is more useful for antennas with thin layers. The inner conductor of the coaxial probe is connected to the radiation patch of the antenna and its outer conductor is connected to the ground plane of the antenna. The antenna illustrated in Fig.1 is feed by 50 Ohm SMA (SubMiniature version A) feed probe. The antenna diameter is calculated using Eq. (1) and Eq. (2). The antenna diameter is calculated using Eq. (1) and Eq. (2).

$$F = \frac{8,791x10^{9}}{f_r\sqrt{\varepsilon_r}} \qquad (1)$$

$$a = \frac{F}{\left\{1+\frac{2h}{\pi\varepsilon_r F\left[ln(\frac{\pi F}{2h})+1,7726\right]^{1/2}}\right\}} \qquad (2)$$

Here, ε_r represents relative permittivity of the substrate, f_r represents resonant frequency, h represents height of the substrate, and a represents radius of the patch.

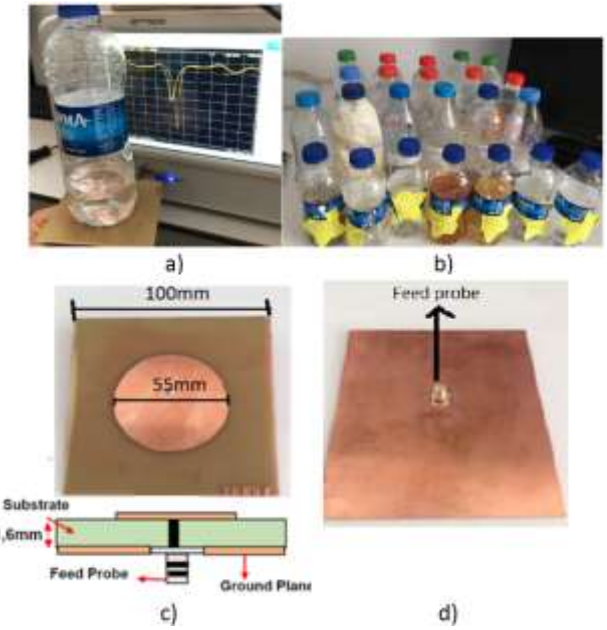


Fig.1. a) Measurement system used in this study b) Samples c) The geometry of the antenna and front view of the antenna d) Back view of the antenna.

A. Support Vector Machines (SVM)

SVM is an algorithm based on statistical learning theory [24]. This algorithm was originally developed for binary classifications [25]. The algorithm is based on the principle of classifying data by finding the best hyperplane that distinguishes the data of a class from those of the other class [26]. A decision function derived from training data is used to find the optimum hyperplane. In a classification problem that can be linearly divided into two classes, k is a training set showing the number of samples, if $x \in$ is an N-dimensional space, $y \in \{-1, +1\}$ is class labels and b is the trend value, then the support vectors are the points that make up the hyperplanes and they are expressed as in Eq. (3). Hyperplane inequalities are given in Eq. (4) and Eq. (5).

$$w.x_i + b = \pm 1 \qquad (3)$$

$$w.x_i + b \geq +1 \quad \text{for each y=+1} \qquad (4)$$

$$w.x_i + b \leq +1 \quad \text{for each y= -1} \qquad (5)$$

where *w* is the normal of the hyperplane and is known as the weight vector [27].

According to the algorithm, the limit of the optimum hyperplane must be maximum. In this case, finding the most suitable hyperplane is possible with the solution of the limited optimization problem given in Eq. (6). Constraints due to this are as shown in Eq. (7) [24]. If this problem is solved by Lagrange equations, Eq. (8) is obtained. For data that can be divided into two classes linearly, the decision function is as given by Eq. (9) [27].

$$min\frac{1}{2}\|w\|^2 \qquad (6)$$

$$y_i(w.x_i + b) - 1 \geq 0 \;\; ve \; y_i \in \{1, -1\} \qquad (7)$$

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{k}\alpha_i y_i(w.x_i + b) + \sum_{i=1}^{k}\alpha_i \qquad (8)$$

$$f(x) = sign\left(\sum_{i=1}^{k}\lambda_i y_i(x.x_i) + b\right) \qquad (9)$$

In some cases, a linear hyperplane that can categorize data cannot be found. In this case, the data in the feature space is moved to a higher dimensional kernel space and a classification is made by finding a hyperplane in the kernel space [28]. Some of the data that cannot be separated linearly remain on the other side of the optimal hyperplane. This problem is solved by defining a positive variable (ξ). The C correction parameter is used to check incorrect classifications. Thus, the problem can be expressed as in Eq. (10). The limitations related to this are as given in Eq. (11).

$$min\left[\frac{\|w\|^2}{2} + C.\sum_{i=1}^{r}\xi_i\right] \qquad (10)$$

$$y_i(w.\phi(x_i) + b) - 1 \geq 1 - \xi_1$$
$$\xi \geq 0 \text{ and } i = 1, \ldots., N \qquad (11)$$

The kernel function used to separate data that cannot be separated linearly is defined mathematically as given by Eq. (12). With the help of this function, data can be classified by making nonlinear transformations. The decision rule for this is as given by Eq. (13) [27].

$$K(x_i, x_j) = \phi(x_i).\phi(x_j) \qquad (12)$$

$$f(x) = sign\left(\sum_{i=1}^{k}\alpha_i\, y_i \phi(x).\phi(x_i) + b\right) \qquad (13)$$

The accuracy of SVM algorithm depends on the selected kernel function [29]. The kernels used in this study are Linear, Polynomial, Radial Basis and Sigmoid. Linear kernel is the simplest kernel function. It is as given in Eq. (14). Polynomial kernel is as given by Eq. (15).

$$\mathrm{K}(\mathrm{x_i}, \mathrm{x_j}) = \mathrm{x_i^T x_j} \qquad (14)$$

$$\mathrm{K}(\mathrm{x_i}, \mathrm{x_j}) = (\mathrm{x_i^T x_j} + \mathrm{d})^{\mathrm{P}} \qquad (15)$$

where, *d* is constant term and *p* is polynomial degree.

Gaussian kernel expressed by Eq. (16) is an example of radial basis function kernel. For this kernel, σ determines the width of the Gaussian kernel and plays a major role for the kernel's performance.

$$\mathrm{K}(\mathrm{x_i}, \mathrm{x_j}) = \exp(-\|\mathrm{x_i} - \mathrm{x_j}\|^2/2\sigma^2 \qquad (16)$$

The Hyperbolic Tangent Kernel expressed by Eq. (17) is also known as Sigmoid Kernel and as Multilayer Perceptron kernel. Here, α represents slope and δ represents intersection constant.

$$\mathrm{K}(\mathrm{X_i}, \mathrm{X_j}) = \tanh(\alpha \mathrm{x_i^T x_j} + \delta) \qquad (17)$$

III. Results

All liquids in the study were measured in pet bottles. 52 liquids were measured, including 23 hazardous liquids and 29 non-hazardous liquids. Liquids that contain alcohol-water solutions with an alcohol content of 70% or more can be hazardous; therefore, these solutions are classified as hazardous liquids [2]. With the proposed measurement system, spectra of liquids were collected in 56 steps in the 1.42-1.53 GHz frequency range and were used as input to SVM algorithm. Liquids used in this study are listed in Table I.

TABLE I
LIQUIDS USED IN THIS STUDY.

Hazardous liquids		**Non-hazardous liquids**		
Ethanol (70,80,90,100)%	Acetone	Peach juice	Vinegar	Turnip juice
Methanol (70,80,90,100)%	Cologne	Shower gel	Shampoo	Champagne
1-propanol (70,80,90,100)%	Toluene	Hair conditioner	Screen cleaning fluid	Tequila
Isopropanol (70,80,90,100)%	Butanol	Lens solution	Buttermilk	Whiskey
Gasoline	Octanol	Ketchup	Apricot juice	Cocoa milk
Thinner		Beer	Liqueur	Hair gel
		Baby food	Water	Liquid soap
		White wine	Red wine	Milk
		Cola	Tea	Gin
		Vodka	Raki	

The amplitude spectra of the aqueous alcohol solutions are given in Fig.2a, the amplitude spectra of the pure hazardous liquids in Fig.2b, and finally the amplitude spectra of the non-hazardous liquids in Fig.2c. Flowchart of the proposed approach is given in Fig. 3.

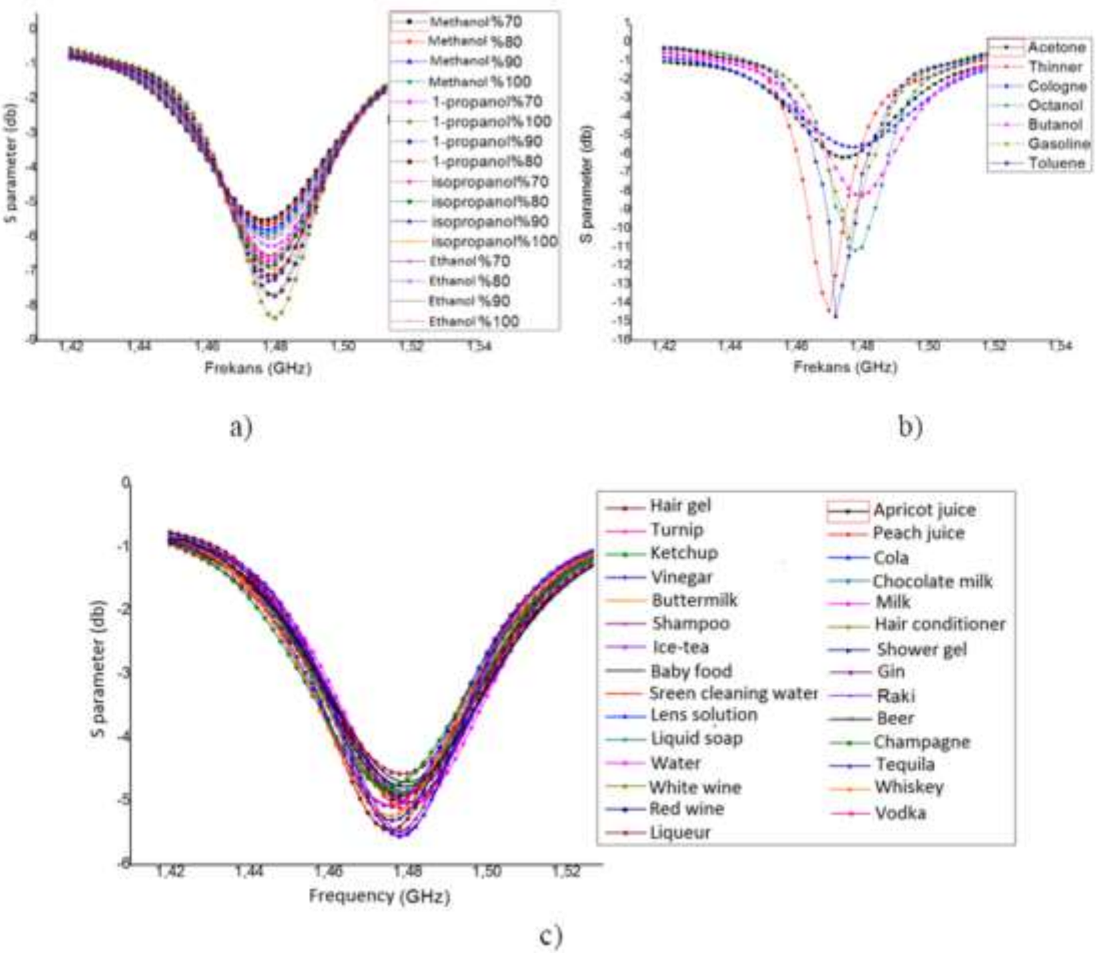


Fig.2. Amplitude spectra of liquids a) Aqueous alcohol solutions b) Hazardous liquids c) Non-hazardous liquids.

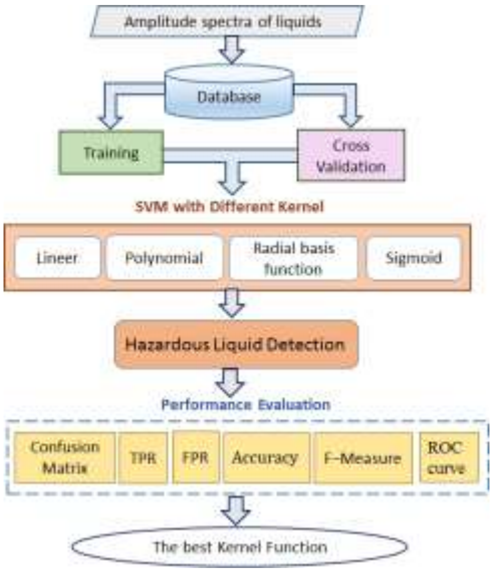


Fig.3. Flowchart of the proposed approach

A. Performance Evaluation

Since kernel functions can affect the performance of the algorithm, to analyze its effects a classification was made using four different kernel functions. K-fold cross validation technique was used in the performance evaluation of this study. In this technique, a dataset is divided into parts and some of it is used as training data and the remaining is used as test data. In this way, it is understood how the model performs on a dataset that it has not seen before. Therefore, a successful algorithm must be able to make an accurate prediction even about a liquid type it does not know at all.

1) Performance metrics

Performance criteria used in this study consists of False Positive Rate (FPR), True Positive Rate (TPR), Confusion matrix, ROC curve, F-Measure, and accuracy. Confusion matrix is a matrix containing information about the actual classes and predicted classes of liquids and is given in Fig.4. The diagonal values of the matrix in green show the number of liquids that the algorithm predicts correctly (True Positive (TP) and True Negative (TN)), and the cells in pink show the number of liquids that it predicts incorrectly (False Negative (FN) and False Positive (FP)). When Fig. 4 is examined, it can be seen that all 52 liquids were correctly classified in both training and cross validation when linear and polynomial functions were preferred. Radial basis function classified 3 hazardous liquids incorrectly. Sigmoid function classified all of the hazardous liquids as non-hazardous.

TPR is the ratio of true positive samples. It is also called recall and is calculated using Eq. (18). FPR is the ratio of false positive samples and is calculated using Eq. (19).

$$TPR = \frac{TP}{TP+FN} \qquad (18)$$

$$FPR = \frac{FP}{TN+FP} \qquad (19)$$

Accuracy shows the overall performance of the model. It is the most popular performance evaluation measure and can be calculated using Eq. (20). F-Measure is a hybrid metric useful for unbalanced classes and is calculated using Eq. (21).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (20)$$

$$F - Measure = \frac{2TP}{2TP+FP+FN} \qquad (21)$$

In a successful classification, F-Measure and TP values are desired to be as close to 1 as possible.

2) Experimental results

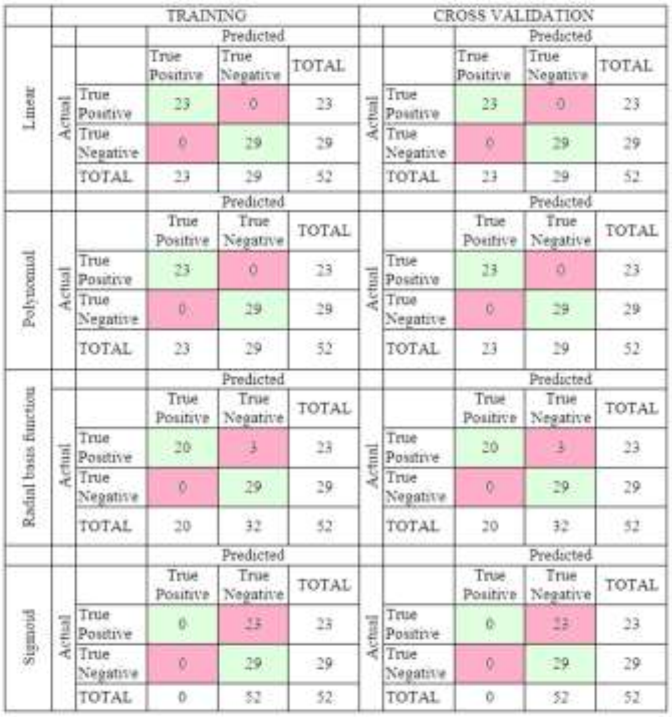
TPR, FPR, F-Measure metrics and accuracy values of all the kernels are listed in Table II. In a ROC curve, FPR is on the X axis and TPR is on the Y axis. As the remaining under the curve increases, the discrimination performance between classes increases. The ROC curves are given in Fig.5 for different kernels.

As it can be seen from the ROC curves of polynomial and linear functions, liquids were correctly classified. The areas under the ROC curves took the value of 1. On the other hand, as it can be seen from the ROC curves of Radial basic function, some misclassifications were done so the ROC area value was calculated as 0.93. Finally, as it can be seen from the ROC curves of Sigmoid function, the function did not classify any liquids correctly, so the ROC area value was calculated as 0.5. This indicates that Sigmoid function is not suitable for liquid classification when SVM algorithm has been preferred.

When Table II is examined, it is seen that the algorithm achieved 100% accuracy by correctly predicting all the liquids when Polynomial and linear kernels were used. Therefore, Kappa, F-Measure and TP values were 1. Moreover, Root Mean Squared Error value of 0 indicates that the algorithm made the classification error-free. When Sigmoid kernel function was used, the algorithm did not accurately predict any hazardous liquid. Therefore, F-Measure value was not computed.

TABLE II
AVERAGE VALUES OF PERFORMANCE METRICS OF DIFFERENT KERNEL FUNCTIONS.

Metric	**Linear**	**Polynomial**	**Radial basis function**	**Sigmoid**
TPR	1	1	0.94	0.55
FPR	0	0	0.07	0.55
F-Measure	1	1	0.94	---
Accuracy (%)	100	100	94.23	55.76



			TRAINING			CROSS VALIDATION		
			Predicted			Predicted		
			True Positive	True Negative	TOTAL	True Positive	True Negative	TOTAL
Linear	Actual	True Positive	23	0	23	23	0	23
		True Negative	0	29	29	0	29	29
		TOTAL	23	29	52	23	29	52
Polynomial	Actual	True Positive	23	0	23	23	0	23
		True Negative	0	29	29	0	29	29
		TOTAL	23	29	52	23	29	52
Radial basis function	Actual	True Positive	20	3	23	20	3	23
		True Negative	0	29	29	0	29	29
		TOTAL	20	32	52	20	32	52
Sigmoid	Actual	True Positive	0	23	23	0	23	23
		True Negative	0	29	29	0	29	29
		TOTAL	0	52	52	0	52	52

Fig.4. Performance of different kernel functions.

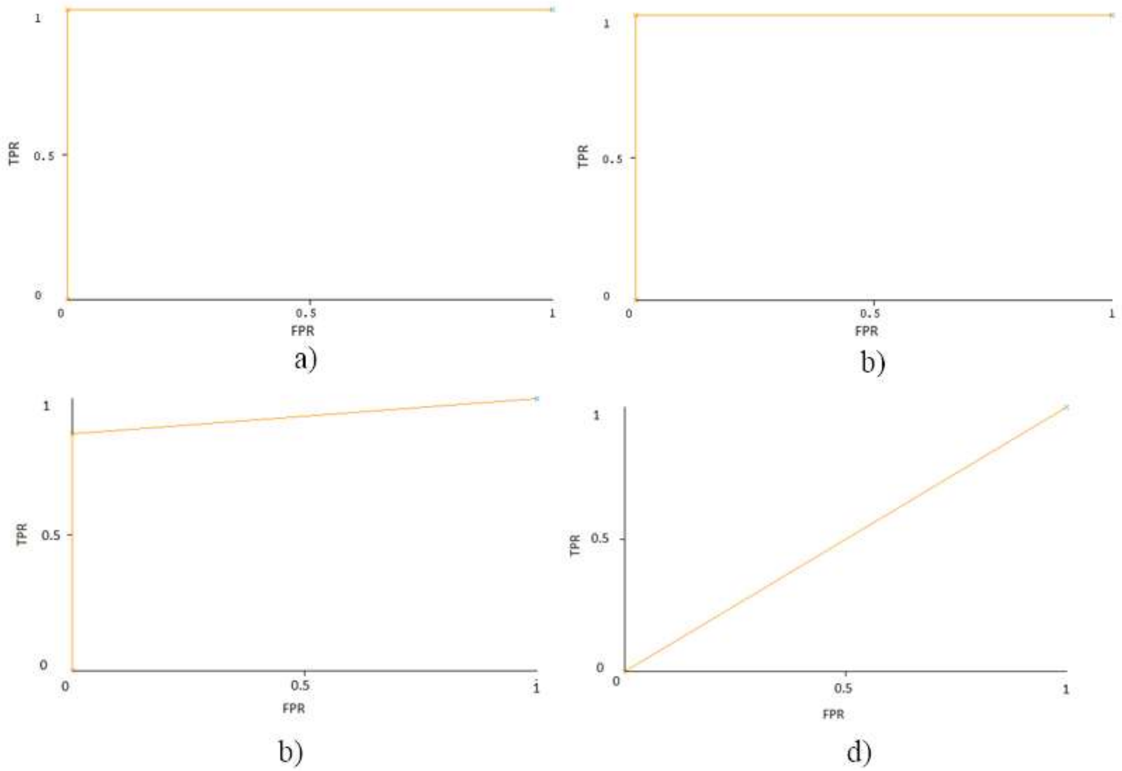


Fig.5. ROC curve of hazardous liquids with different kernel functions a)Linear b) Polynomial c) Radial basis d) Sigmoid.

IV. Conclusion

In this study, a system that can be used for non-contact detection of hazardous liquids is proposed. With this system, hazardous liquids can be evaluated independently of operator interpretation. Therefore, the obtained result does not vary from person to person. In addition, as the measuring system can scan liquids with a closed container, it prevents the operator from experiencing health problems caused by smelling these hazardous liquids and touching them.

When the measurement system and the embedded SVM algorithm are used, hazardous liquids can be detected quickly and with high accuracy. Kernel functions (Linear, Polynomial, Sigmoid or Radial) can be used to increase the success of the SVM algorithm. As a result of the classification made when Linear and Polynomial functions were used, the TPR value was 1 and the FPR value was 0. In addition, the accuracy rate was 100%. These values prove that the algorithm can classify all liquids correctly when Linear and Polynomial functions have been used. The confusion matrices and ROC curves presented in this paper support this conclusion. Consequently, it is recommended to use Polynomial or Linear kernel functions in the proposed system.

References

[1] W. Luo, Z. Zhang, H. Liu, C. Zhang. “Terahertz reflection time-domain spectroscopy for measuring alcohol concentration.” Infrared, Millimeter- Wave, and Terahertz Technologies V, International Society for Optics and Photonics, 2018, pp. 1082615. doi:10.1117/12.2500966

[2] X. Tan, S. Huang, Y. Zhong, H. Yuan, Y. Zhou, Q. Xiao, L. Guo, S. Tang, Z. Yang, C. Qi. “Detection and identification of flammable and explosive liquids using THz time-domain spectroscopy with principal component analysis algorithm.” 2017 10th UK-Europe-China Workshop on Millimetre Waves and Terahertz Technologies UCMMT , IEEE, 2017, pp. 1-4. doi:10.1109/UCMMT.2017.8068488

[3] X. Tan, S. Tang, Z. Yang, J. Xie, J. Tang, F. Xie, C. Qi. “ Detection and identification of liquids using reflection THz time-domain spectroscopy with principal component analysis and support vector machine algorithm.” International Symposium on Ultrafast Phenomena and Terahertz Waves, Optical Society of America, 2018, pp. WI27. doi:10.1364/ISUPTW.2018.WI27

[4] W. Zhang, Y. Tang, A. Shi, L. Bao, Y. Shen, R. Shen, Y. Ye. “ Recent developments in spectroscopic techniques for the detection of explosives.” Materials, 11 2018 1364. doi:10.3390/ma11081364

[5] M.F. Isaac-Lam. “ Incorporation of Benchtop NMR Spectrometer into the Organic Chemistry Laboratory: Analysis of an Unknown Liquid.” Journal of Chemical Education, 97 2020, pp. 2036-2040. doi:10.1021/acs.jchemed.9b00787

[6] E. Gudmundson, A. Jakobsson, I.J. Poplett, J.A. Smith. “ Detection and classification of liquid explosives using NMR.” 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2009, pp. 3053-3056.

[7] M.L. Ramírez-Cedeño, W. Ortiz- Rivera, L.C. Pacheco-Londoño, S.P. Hernández-Rivera. “Remote detection of hazardous liquids concealed in glass and plastic containers.” IEEE Sensors Journal, 10, 2010,pp 693-698. doi:10.1109/JSEN.2009.2036373

[8] P. Orachorn, N. Chankow, S. Srisatit. “Development of technique for screening liquids in glass bottle using low energy X-ray transmission.” RMUTT Research Journal Rajamangala University of Technology Thanyaburi, 16, 2017, pp 20-26.

[9] S.I.Y. Al-Mously. " A modified complex permittivity measurement technique at microwave frequency." International Journal of New Computer Architectures and Their Applications, 2, 2012, pp 389-402.
[10] Z. Li, A. Haigh, C. Soutis, A. Gibson, R. Sloan. "A simulation-assisted non- destructive approach for permittivity measurement using an open-ended microwave Waveguide." Journal of Nondestructive Evaluation, 37, 2018, 39. doi:10.1007/s10921-018-0493-1
[11] R.V. Shinde, A.R. Deshmukh, S.A. Ingole, A.C. Kumbharkhane. "Dielectric spectroscopy and hydrogen bonding studies of 1-chloropropane– ethanol mixture using TDR technique, Journal of Advanced Dielectrics." 9, 2019, 1950018. doi:10.1142/S2010135X19500188
[12] V. Gaiduk, S. Nikitov. "Possible mechanisms of dielectric relaxation of liquid water and calculation of the temperature dependence of the complex permittivity of water." Optics and Spectroscopy, 98, 2005, pp 919-933. doi:10.1134/1.1953988
[13] V. Gaĭduk. "Relations between the association of liquid water molecules and the dielectric and raman spectra of H 2 O." Optics and Spectroscopy, 106, 2009, 24-42. doi:10.1134/S0030400X09010044
[14] G. Gennarelli, S. Romeo, M.R. Scarfì, F. Soldovieri. "A microwave resonant sensor for concentration measurements of liquid solutions." IEEE Sensors Journal, 13, 2013 pp 1857-1864. doi:10.1109/JSEN.2013.2244035
[15] M.A. Sairin, N.H. Abd Latiff, S. Abd Aziz, F.Z. Rokhani. "Distinguishing edible oil using dielectric spectroscopy at microwave frequencies of 8.2–12.1 GHz." 2016 10th International Conference o n Sensing Technology ICST , IEEE, 2016, pp. 1-4. doi:10.1109/ICSensT.2016.7796333
[16] L. Zhang, M.A. Schultz, R. Cash, D.M. Barrett, M.J. McCarthy. "Determination of quality parameters of t omato paste using guided microwave spectroscopy." Food control, 40, 2014, pp 214-223. doi:10.1016/j.foodcont.2013.12.008
[17] A.V. Yurchenko, A. Novikov, M.V. Kitaeva. "A resonator microwave sensor for measuring the parameters of Solar-quality silicon." Russian Journal of Nondestructive Testing, 48, 2012, pp 109-114. doi:10.1134/S1061830912020118
[18] A. La Gioia, E. Porter, I. Merunka, A. Shahzad, S. Salahuddin, M. Jones, M. O'Halloran. "Open-ended coaxial probe technique for dielectric measurement of biological tissues: Challenges and common practices." Diagnostics, 8, 2018, 40. doi:10.3390/diagnostics8020040
[19] P. Hamsagayathri, P. Sampath. "Microwave Breast Cancer Screening for Women Welfare, Indian Journal of Public Health Research & Development." 8, 2017, pp 115-121.
[20] A. Gupta & V.R. Kumar, "Machine Learning Technology Using Thick Film Gas Sensor Toxic Liquid Detection For Industrial IOT Application". In 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020, pp. 1-6. IEEE.
[21] A. Gupta, S.K. Dargar & M. Sabir. "Performance analysis of 2% Fe 2 O 3 Doped Thick-film Gas Sensor in Toxic Liquid Detection Using Machine Learning Techniques". In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022, pp. 689-693ş, IEEE.
[22] R. Wellock, A.D. Walmsley. "Applications of microwave spectroscopy in process analysis." Spectroscopy Europe, 16, 2004, pp 23- 26.
[23] P. Singh, S. Bhamidipati, R. Singh, R. Smith, P. Nelson. "Evaluation of in-line sensors for prediction of soluble and total solids/moisture in continuous processing of fruit juices." Food Control, 7, 1996, pp 141-148. doi:10.1016/0956-7135 96 00020-5
[24] C. Cortes, V. Vapnik, Support-vector networks. "Machine learning." 20, 1995, pp 273-297.
[25] G.M. Foody, A. Mathur. "Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification." Remote Sensing of Environment, 93, 2004, pp 107-117.
[26] D.D. Gutierrez. "Machine learning and data science: an introduction to statistical learning methods with R." Technics Publications, 2015. doi:10.1109/ICASSP.2009.4960268
[27] E.E. Osuna. "Support vector machines: Training and applications." Massachusetts Institute of Technology, 1998.
[28] Z. Liu, M.J. Zuo, X. Zhao, H. Xu, An. "Analytical Approach to Fast Parameter Selection of Gaussian RBF Kernel for Support Vector Machine." J. Inf. Sci. Eng., 31, 2015, pp 691-710.
[29] T. Kavzoglu, I. Colkesen. "A kernel functions analysis for support vector machines for land cover classification, International Journal of Applied Earth Observation and Geoinformation." 11, 2009, pp 352-359. doi:10.1016/j.jag.2009.06.002

BIOGRAPHIES

EBRU EFEOGLU is currently an Assistant Professor at Kütahya Dumlupinar University Software Department. She received her B.S. degree in Geophysics Engineering and Management Information. She received her Ph.D. degree in Computer Engineering from Trakya University, Turkey in 2021. She has authored several papers in international conference proceedings and SCI-Expanded journals. Her research interests include machine learning and data mining, and their applications in various research domains.
Ebru Efeoglu.

GURKAN TUNA is currently a Professor at the Department of Computer Programming at Trakya University, Turkey. He is also the head of the graduate program of Mechatronics Engineering at the same university. His current research interests include wireless networks, wireless sensor networks, multi-sensor fusion, and smart cities.

Detection of TrickBot and Emotet Banking Trojans with Machine Learning

Ruveyda Celik and Ali Gezer

Abstract— **Internet banking is getting more popular with the increasing number and demand of online banking customers. Almost all transactions that could be performed in bank branches could also be realized through internet banking. Internet banking, which has become widespread with the increasing use of the Internet, has also led to an increase in cases of financial fraud. This has made the protection of personal data and the security of banking services more important than ever. It is very important for institutions and organizations providing online banking services to take security measures in their systems. Cybercriminals target internet users with methods such as malware infection, botnets, spam, phishing, identity theft, and social engineering that they use and develop every day. Therefore, there are always potential risks in using internet banking. Banking viruses commonly used by cybercriminals today are TrickBot and Emotet. Nowadays TrickBot and Emotet are popular banking trojans which gives hard times for onlıne banking customers. Their primary goal is to steal user's banking and personal information. In this study, we will investigate the behavior analysis and new tricks of TrickBot and Emotet banking viruses, which use different methods to compromise the security of online banking customers. We benefited WEKA program to detect these banking viruses. In addition to this, we also focused on the detection of TrickBot and Emotet Banking viruses with using Random Tree, J48, Naive Bayes, SMO Techniques.**

Index Terms— ***Banking Trojan, Emotet, Machine Learning Methods, Malware Analysis, TrickBot, Web Injections***

I. Introduction

BANKING TROJANS are viruses that pretend to be a legitimate program or file, infiltrate computers and perform harmful actions. Although no one wants to be exposed to cyberattacks, millions of people become victim of attackers each year. In addition, banking viruses can create a backdoor that can copy the credentials of a bank customer by imitating the login web page of financial institutions.

TrickBot and Emotet are popular banking trojans that make such transactions from online banking and finance sites to attackers digital systems. Emotet is a Trojan that is primarily spread through spam emails (malspam). The malware may infect either via malicious script, macro-enabled document files, or malicious links. Emotet emails may contain familiar branding which designed to look like a legitimate email. Emotet may try to persuade users to click the malicious files by using tempting language such as "Your Invoice," "Payment Details," or possibly an upcoming shipment from well-known parcel companies [1].

Emotet has gone through a few iterations. Early versions arrived as malicious JavaScript file. Later versions evolved to use macro-enabled documents to retrieve the virus payload from command and control (C&C) servers.

Emotet uses a number of tricks to try and prevent detection and analysis. Notably, Emotet knows if it's running inside a virtual machine (VM) and will lay dormant if it detects a sandbox environment, which is a tool cybersecurity researchers use to observe malware within a safe, controlled space [2].

Emotet also uses C&C servers to receive updates. This works in the same way as the operating system updates on your PC and can happen seamlessly and without any outward signs. This allows the attackers to install updated versions of the software, install additional malware such as other banking trojans, or to act as a dumping ground for stolen information such as financial credentials, usernames, passwords, and email addresses [3].

TrickBot was created to steal users' banking information. When Malwarebytes researchers first discovered TrickBot in 2016, they thought it was an ordinary identity theft purpose malware. But TrickBot targeted financial services and users for their banking data. It has also exploited other malwares to achieve its goals [4].TrickBot has the reputation of being the successor to another credential thief, Dyreza, who first appeared in 2014. TrickBot shared similarities with Dyreza, such as certain variables with similar values and the way that the functioning of command and control (C&C) servers. This has led many researchers to believe that the person or group that created Dyreza also created TrickBot [5].

In 2017, developers added a worm module to TrickBot, which we believe was inspired by successful ransomware campaigns with worm-like capabilities such as WannaCry and EternalPetya [6]. The developers also added a module for collecting Outlook credentials. The reason for adding this module is that hundreds of organizations and millions of people around the world often use this web mail service. The range of data TrickBot plays has also expanded. These are: cookies, browsing history, URLs visited, Flash LSO (Local Shared Objects) and many more. Although these modules were new at that time, they weren't coded well enough.

In 2018, TrickBot continued to exploit the SMB

RÜVEYDA ÇELİK, is with Department of Electrical and Electronics Engineering University of Kayseri University, Kayseri, Turkey, (e-mail: ruveydacelik38@hotmail.com).

https://orcid.org/0000-0003-4821-4633

ALİ GEZER, is with Department of Cyber Security Application and Research Center University of Kayseri, Kayseri, Turkey, (e-mail: agezer@kayseri.edu.tr).

https://orcid.org/0000-0001-8265-1736

Manuscript received Dec 1, 2021; accepted Jul 4, 2022.
DOI: 10.17694/bajece.1031021

vulnerability. It was also equipped with the module that disables Windows Defender's real-time monitoring using a PowerShell command. While it had also updated its encryption algorithm, the rest of its module function stayed the same. TrickBot developers also started securing their code from being taken apart by security researchers via incorporating obfuscation elements [7]. At the end of the year, TrickBot was ranked as the top threat against businesses, and has overtaken Emotet. TrickBot developers made some changes to the Trojan in 2019. Specifically, they made changes to the way that webinject feature works against the some US-based mobile carriers.

Recently, researchers have noted an improvement in this Trojan's evasion method. Mworm, the module responsible for spreading a copy of itself, was replaced by a new module called Nworm. This new module alters TrickBot's HTTP traffic, allowing it to run from memory after infecting a domain controller. This ensures that TrickBot doesn't leave any traces of infection on affected machines.

These banking viruses have allowed them to install updated versions of the software, install additional malware such as other banking Trojans, or act as intermediaries for stolen information such as financial credentials, usernames, passwords, and email addresses [8]. The chain of infection diagram for banking malicious viruses for Emotet and TrickBot is shown in Figure 1.

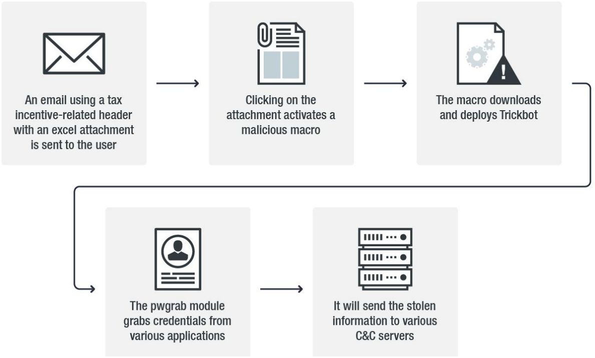
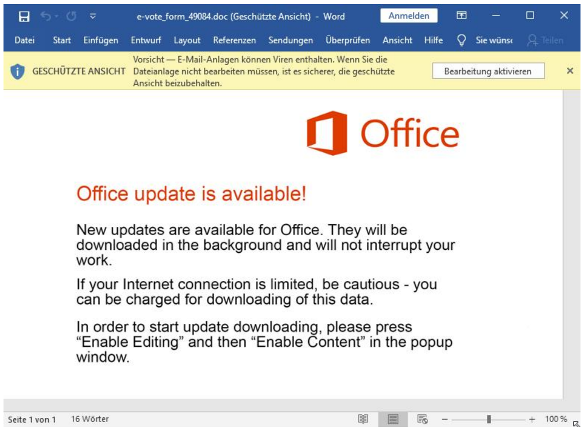

An email using a tax incentive-related header with an excel attachment is sent to the user
Clicking on the attachment activates a malicious macro
The macro downloads and deploys Trickbot
The pwgrab module grabs credentials from various applications
It will send the stolen information to various C&C servers

Fig.1. Infection chain for Emotet and TrickBot

II. PROSED METHOD

A. Data Collection Through Static And Dynamic Analysis

Static and dynamic analyzes could be performed to reveal the signatures of malicious softwares. Via determining network flows, virus detection was investigated in virtual machines infected with TrickBot and Emotet. When the TrickBot and Emotet trojans infect a system, their first action is to identify their victims. It performs a network activity via e-mail or HTTP request to some service sites, websites of targeted banks or websites where users can access their personal data.



Vorsicht — E-Mail-Anlagen können Viren enthalten. Wenn Sie die Dateianlage nicht bearbeiten müssen, ist es sicherer, die geschützte Ansicht beizubehalten.

Office update is available!

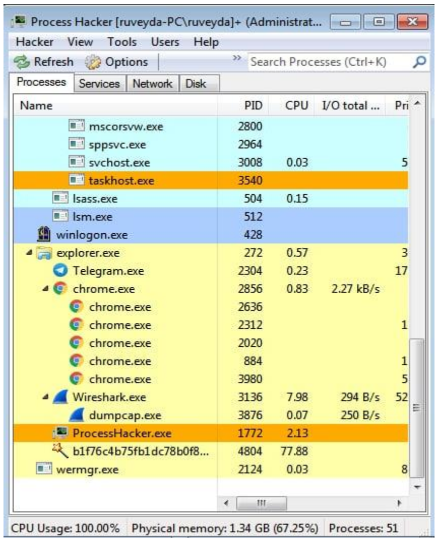
New updates are available for Office. They will be downloaded in the background and will not interrupt your work.

If your Internet connection is limited, be cautious - you can be charged for downloading of this data.

In order to start update downloading, please press “Enable Editing” and then “Enable Content” in the popup window.

Fig.2. Suspicious e-mail content

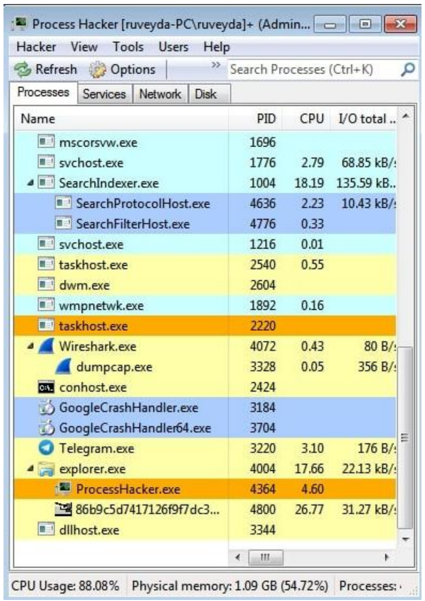
When the infected file in the email is executed the compromised system tries to connect to one of Command and Control servers of these two currently active banking viruses. Some server IPs are encoded into the malware's binary. After the connection is established, TrickBot and Emotet viruses try to download the encrypted file modules. They try to access new IPs by leaking stolen data to Downloaders. We can see the IPs that these viruses interact with using Process Hacker and Wireshark programs.


Process Hacker [ruveyda-PC\ruveyda]+ (Administrat...
Hacker
View
Tools
Users
Help
Refresh
Options
Search Processes (Ctrl+K)
Processes
Services
Network
Disk

Name	PID	CPU	I/O total ...	Pri
mscorsvw.exe	2800			
sppsvc.exe	2964			
svchost.exe	3008	0.03		5
taskhost.exe	3540			
lsass.exe	504	0.15		
lsm.exe	512			
winlogon.exe	428			
explorer.exe	272	0.57		3
Telegram.exe	2304	0.23		17
chrome.exe	2856	0.83	2.27 kB/s	
chrome.exe	2636			
chrome.exe	2312			1
chrome.exe	2020			
chrome.exe	884			1
chrome.exe	3980			5
Wireshark.exe	3136	7.98	294 B/s	52
dumpcap.exe	3876	0.07	250 B/s	
ProcessHacker.exe	1772	2.13		
b1f76c4b75fb1dc78b0f8...	4804	77.88		
wermgr.exe	2124	0.03		8

CPU Usage: 100.00%   Physical memory: 1.34 GB (67.25%)   Processes: 51

Fig.3. Process Hacker output of a computer infected with Emotet


Process Hacker [ruveyda-PC\ruveyda]+ (Admin...

Hacker View Tools Users Help
Refresh Options
Search Processes (Ctrl+K)
Processes Services Network Disk

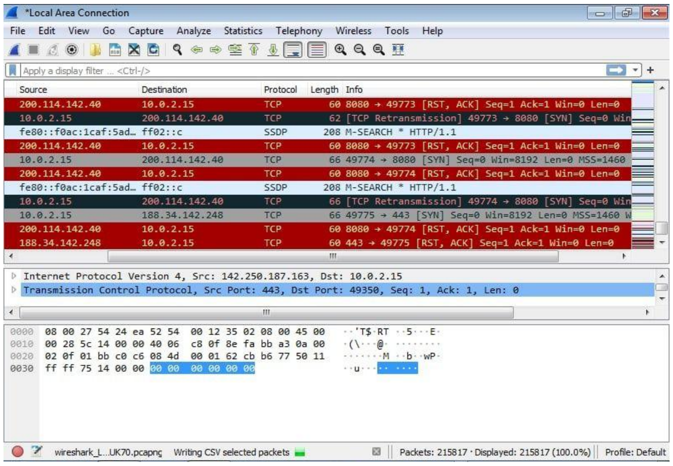
Name	PID	CPU	I/O total ..
mscorsvw.exe	1696		
svchost.exe	1776	2.79	68.85 kB/:
SearchIndexer.exe	1004	18.19	135.59 kB..
SearchProtocolHost.exe	4636	2.23	10.43 kB/:
SearchFilterHost.exe	4776	0.33	
svchost.exe	1216	0.01	
taskhost.exe	2540	0.55	
dwm.exe	2604		
wmpnetwk.exe	1892	0.16	
taskhost.exe	2220		
Wireshark.exe	4072	0.43	80 B/:
dumpcap.exe	3328	0.05	356 B/:
conhost.exe	2424		
GoogleCrashHandler.exe	3184		
GoogleCrashHandler64.exe	3704		
Telegram.exe	3220	3.10	176 B/:
explorer.exe	4004	17.66	22.13 kB/:
ProcessHacker.exe	4364	4.60	
86b9c5d7417126f9f7dc3...	4800	26.77	31.27 kB/:
dllhost.exe	3344		

CPU Usage: 88.08% Physical memory: 1.09 GB (54.72%) Processes:

Fig.4. Process Hacker output of a computer infected with TrickBot

Process Hacker is an open source tool that allows you to see what processes are running on a device, identify programs consuming CPU resources, and identify network connections associated with a process. Such features make Process Hacker an ideal tool for monitoring malware on a device. Being able to see what processes are spawned identify network connections and interesting threads could give us valuable indicators of danger (IOCs) [9].

IP addresses and malicious domain names are valuable indicators in incident response. Using Process Hacker is helpful to gather such information which also compromised hosts can be identified in the network.


*Local Area Connection

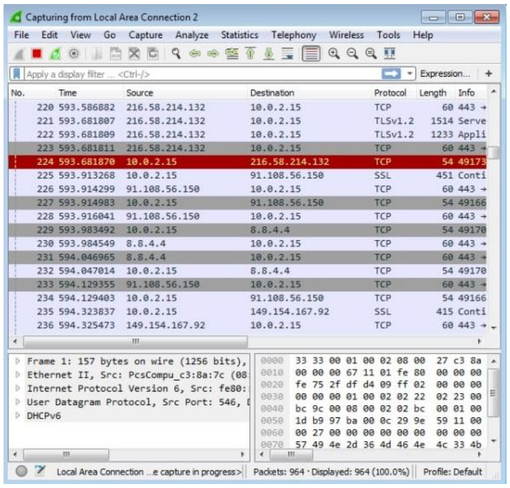
File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help
Apply a display filter ... <Ctrl-/>

```
Source                Destination     Protocol Length Info
200.114.142.40        10.0.2.15       TCP       60 8080 → 49773 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
10.0.2.15             200.114.142.40  TCP       62 [TCP Retransmission] 49773 → 8080 [SYN] Seq=0 Win
fe80::f0ac:1caf:5ad… ff02::c         SSDP     208 M-SEARCH * HTTP/1.1
200.114.142.40        10.0.2.15       TCP       60 8080 → 49773 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
10.0.2.15             200.114.142.40  TCP       66 49774 → 8080 [SYN] Seq=0 Win=8192 Len=0 MSS=1460
200.114.142.40        10.0.2.15       TCP       60 8080 → 49774 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
fe80::f0ac:1caf:5ad… ff02::c         SSDP     208 M-SEARCH * HTTP/1.1
10.0.2.15             200.114.142.40  TCP       66 [TCP Retransmission] 49774 → 8080 [SYN] Seq=0 Win
10.0.2.15             188.34.142.248  TCP       66 49775 → 443 [SYN] Seq=0 Win=8192 Len=0 MSS=1460 W
200.114.142.40        10.0.2.15       TCP       60 8080 → 49774 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
188.34.142.248        10.0.2.15       TCP       60 443 → 49775 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0

▷ Internet Protocol Version 4, Src: 142.250.187.163, Dst: 10.0.2.15
▷ Transmission Control Protocol, Src Port: 443, Dst Port: 49350, Seq: 1, Ack: 1, Len: 0

0000  08 00 27 54 24 ea 52 54  00 12 35 02 08 00 45 00   ··'T$·RT ··5···E·
0010  00 28 5c 14 00 00 40 06  c8 0f 8e fa bb a3 0a 00   ·(\···@· ········
0020  02 0f 01 bb c0 c6 08 4d  00 01 62 cb b6 77 50 11   ·······M ··b··wP·
0030  ff ff 75 14 00 00 00 00  00 00 00 00               ··u··· ·· ····
```

wireshark_L...UK70.pcapng Writing CSV selected packets | Packets: 215817 · Displayed: 215817 (100.0%) | Profile: Default

Fig.5. Wireshark output of a computer infected with TrickBot

Wireshark is a packet sniffer and protocol analysis tool. It captures network traffic in the local network and stores this data for offline analysis. Wireshark captures network traffic from Ethernet, Bluetooth, Wireless (IEEE.802.11), Token Ring, Frame Relay connections and more. Wireshark lets you filter the log before capture starts or during analysis so you can narrow down what you're looking for. For example, you can set a filter to see TCP traffic between two IP addresses. You can set it to show you only packets sent from a computer. The powerful filtering mechanisms in Wireshark is one of the main reasons it has become the standard tool for packet analysis [10].


Capturing from Local Area Connection 2

File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help
Apply a display filter ... <Ctrl-/>
Expression... +

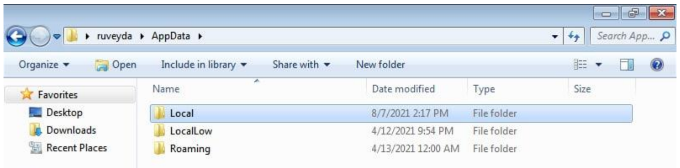
```
No.   Time        Source           Destination      Protocol Length Info
  220 593.586882  216.58.214.132   10.0.2.15        TCP        60 443 →
  221 593.681807  216.58.214.132   10.0.2.15        TLSv1.2  1514 Serve
  222 593.681809  216.58.214.132   10.0.2.15        TLSv1.2  1233 Appli
  223 593.681811  216.58.214.132   10.0.2.15        TCP        60 443 →
  224 593.681870  10.0.2.15        216.58.214.132   TCP        54 49173
  225 593.913268  10.0.2.15        91.108.56.150    SSL       451 Conti
  226 593.914299  91.108.56.150    10.0.2.15        TCP        60 443 →
  227 593.914983  10.0.2.15        91.108.56.150    TCP        54 49166
  228 593.916041  91.108.56.150    10.0.2.15        TCP        60 443 →
  229 593.983492  10.0.2.15        8.8.4.4          TCP        54 49170
  230 593.984549  8.8.4.4          10.0.2.15        TCP        60 443 →
  231 594.046965  8.8.4.4          10.0.2.15        TCP        60 443 →
  232 594.047014  10.0.2.15        8.8.4.4          TCP        54 49170
  233 594.129355  91.108.56.150    10.0.2.15        TCP        60 443 →
  234 594.129403  10.0.2.15        91.108.56.150    TCP        54 49166
  235 594.323837  10.0.2.15        149.154.167.92   SSL       415 Conti
  236 594.325473  149.154.167.92   10.0.2.15        TCP        60 443 →

▷ Frame 1: 157 bytes on wire (1256 bits),   0000  33 33 00 01 00 02 08 00  27 c3 8a
▷ Ethernet II, Src: PcsCompu_c3:8a:7c (08   0010  00 00 00 67 11 01 fe 80  00 00 00
▷ Internet Protocol Version 6, Src: fe80:   0020  fe 75 2f df d4 09 ff 02  00 00 00
▷ User Datagram Protocol, Src Port: 546, [   0030  00 00 00 01 00 02 02 22  02 23 00
▷ DHCPv6                                   0040  bc 9c 00 08 00 02 02 bc  00 01 00
                                           0050  1d b9 97 ba 00 0c 29 9e  59 11 00
                                           0060  00 27 00 00 00 00 00 00  00 00 00
                                           0070  57 49 4e 2d 36 4d 46 4e  4c 33 4b
```

Local Area Connection ...e capture in progress> | Packets: 964 · Displayed: 964 (100.0%) | Profile: Default

Fig.6. Wireshark output of a computer infected with Emotet

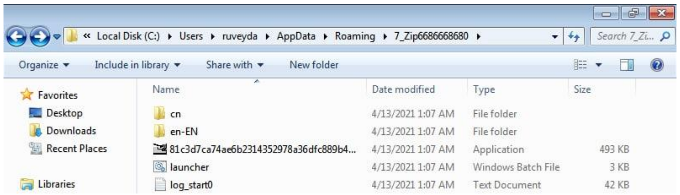
When we look at network traffic, the compromised system tries to make a connection to one of the C&C servers of TrickBot and Emotet. Some server IPs are encoded into the malware's binary. After the connection is established, TrickBot and Emotet viruses try to download the encrypted file modules. It tries to access new IP's by leaking stolen data.

TrickBot and Emotet virus download files in AppData folder. While TrickBot spawns in Roaming, Emotet spawns in Local directory.



Name	Date modified	Type	Size
Local	8/7/2021 2:17 PM	File folder	
LocalLow	4/12/2021 9:54 PM	File folder	
Roaming	4/13/2021 12:00 AM	File folder	

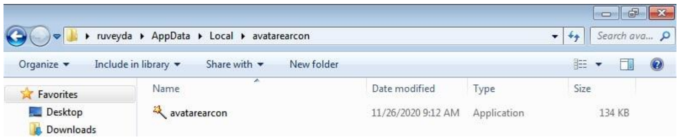
Fig.7. Directories that TrickBot and Emotet cloned themselves


« Local Disk (C:) ▸ Users ▸ ruveyda ▸ AppData ▸ Roaming ▸ 7_Zip6686668680 ▸
Search 7_Zi...
Organize ▾ Include in library ▾ Share with ▾ New folder
Favorites
Desktop
Downloads
Recent Places
Libraries

Name	Date modified	Type	Size
cn	4/13/2021 1:07 AM	File folder	
en-EN	4/13/2021 1:07 AM	File folder	
81c3d7ca74ae6b2314352978a36dfc889b4...	4/13/2021 1:07 AM	Application	493 KB
launcher	4/13/2021 1:07 AM	Windows Batch File	3 KB
log_start0	4/13/2021 1:07 AM	Text Document	42 KB

Figure.8. Created Files after TrickBot infection

TrickBot cloned itself in C:\Users\*\AppData\Roaming\7_Zip6686668680) copied itself and downloaded different files for different purposes.


ruveyda ▸ AppData ▸ Local ▸ avatarearcon
Search ava...
Organize ▾ Include in library ▾ Share with ▾ New folder
Favorites
Desktop
Downloads

Name	Date modified	Type	Size
avatarearcon	11/26/2020 9:12 AM	Application	134 KB

Figure.9. Created Files after Emotet Infection

Emotet cloned itself in

C:\Users\*\AppData\Local\avatarearcom and downloaded files to realized its purposes.

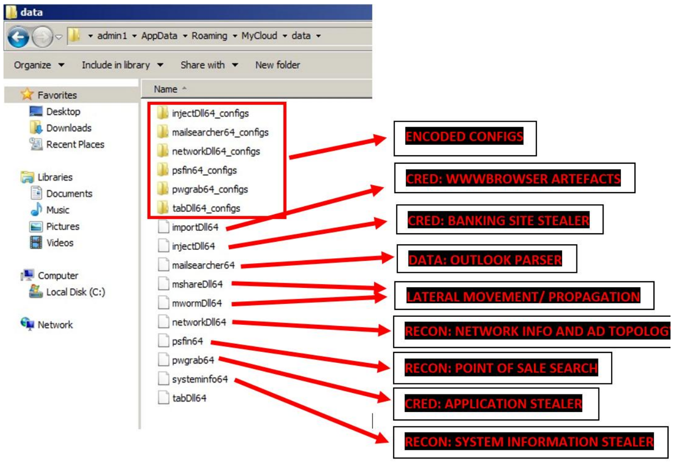


Fig.10. Created files and downloads by Emotet and TrickBot infection

Emotet and TrickBot download files for some specific purposes. Each of these files has different properties and purposes. The most important of these are as follows:

▪ TrickBot and Emotet modules are delivered as Dynamic Link Libraries (DLLs) loaders.

▪ Mainly TrickBot and Emotet have two core modules, Injectdll and systeminfo.

▪ Injectdll module is used to target banking and financial data, monitors banking website activity and uses web injections to steal financial information. Systeminfo is used to fingerprint the infected system specifications [11].

▪ Besides the above features, other files TrickBot and Emotet download are as follows:

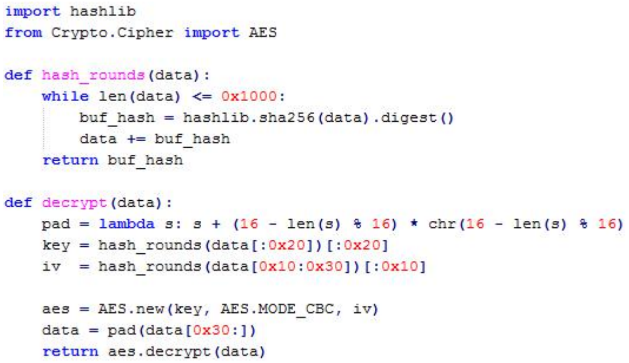
- **ModuleDll/ImportDll:** Collects browser data (eg cookies and browser configurations).
- **Dinj:** File contains bank information; uses server-side web injections.
- **Dpost:** Most of the data leaked by TrickBot is sent to the dpost IP address.
- **Sinj:** Keeps information about targeted online banks; Uses redirect attacks (fake web injections) to leak financial data
- **DomainDll:** Uses LDAP to collect credentials and configuration data. Domain controller by accessing shared SYSVOL files.
- **OutlookDll:** Harvests saved Microsoft Outlook credentials by querying several registry keys.
- **SqulDll:** Force enables WDigest authentication and utilizes Mimikatz to scrape credentials from LSASS.exe. The worming modules use these credentials to spread TrickBot and Emotet laterally across networks.
- **NetworkDll, wormDll, shareDll:** Used for network reconnaissance and lateral movement.
- **RdpScanDll:** Bruteforces RDP for a specific list of victims.

TrickBot banking trojan uses a domain creation algorithm to communicate with its servers. Once infected, the trojan starts executing DNS queries for the created domains. Another popular banking trojan, Emotet, which exhibits a completely different network communication model, sets up a local proxy and routes internet traffic through the proxy server.

Modules can be downloaded from one of TrickBot's or Emotet's C2s using simple GET requests such as https://<CC_IP>:<CC_PORT>/<gtag>/<bot_ID>/5/<module_name>/. Although module names are case sensitive and we define 32-bit modules, in most cases 64-bit versions can be downloaded by typing '64' instead of '32' in the module name. In most cases, valid <gtag> and <bot_ID> values are not required for successful download. Files which are encrypted could be decrypted with the following Python script [12].

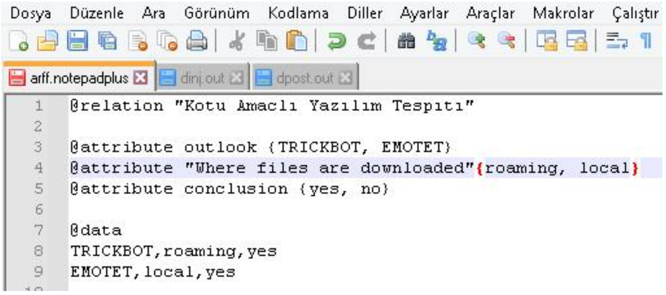
TABLE I
PROPERTIES OF FILES DOWNLOADED TICKBOT AND EMOTET

Name	Function
importDll64	Browser data stealer module
injectDll64	Handles web-injects, including support for several hundred banking/financial sites
mailsearcher64	Recon module parses specific file types for “of interest” data
mshareDll64	Lateral movement / enumeration module via LDAP and SMB exploitation. Mshare and mworm modules work in cooperation
mwormDll64 mshareDll	Lateral movement / enumeration module via LDAP and SMB exploitation. Mshare and mworm modules work in cooperation
networkDll64	Recon module queries network specific environmental data
psfin64	Point-of-sale recon module
pwgrab64	Steals credentials, autofill data, history, and other information from browsers as well as several software applications.
systeminfo64	Recon module. Provides system-specific information and data to the C2
tabDll64	Credential theft module. Sometimes contains additional lateral movement code. Uses the EternalRomance exploit (CVE-2017-0147) to spread via SMBv1.



```python
import hashlib
from Crypto.Cipher import AES

def hash_rounds(data):
    while len(data) <= 0x1000:
        buf_hash = hashlib.sha256(data).digest()
        data += buf_hash
    return buf_hash

def decrypt(data):
    pad = lambda s: s + (16 - len(s) % 16) * chr(16 - len(s) % 16)
    key = hash_rounds(data[:0x20])[:0x20]
    iv  = hash_rounds(data[0x10:0x30])[:0x10]

    aes = AES.new(key, AES.MODE_CBC, iv)
    data = pad(data[0x30:])
    return aes.decrypt(data)
```

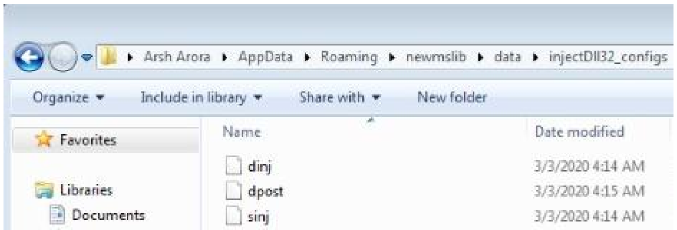
Fig.11. Python script with module decrypt routine



Dosya Düzenle Ara Görünüm Kodlama Diller Ayarlar Araçlar Makrolar Çalıştır
arff.notepadplus
dinj.out
dpost.out

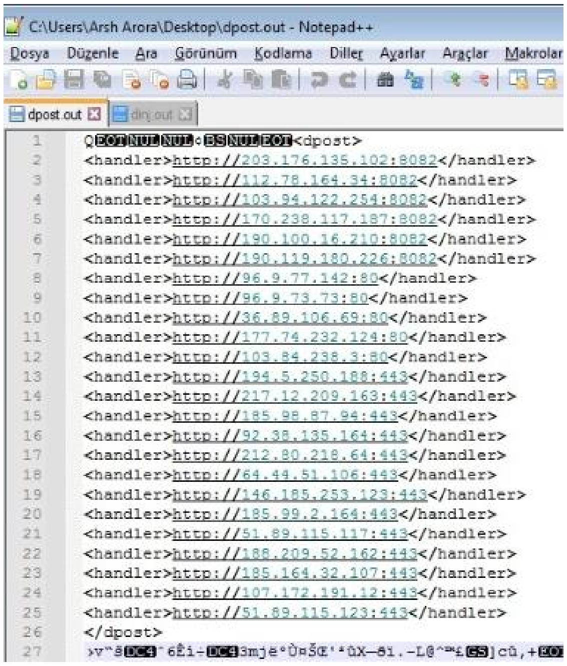
```
@relation "Kotu Amaclı Yazılım Tespitı"

@attribute outlook {TRICKBOT, EMOTET}
@attribute "Where files are downloaded"{roaming, local}
@attribute conclusion {yes, no}

@data
TRICKBOT,roaming,yes
EMOTET,local,yes
```

Fig.12. Notepadplus arff file created as a result of Python password analysis



Name	Date modified
dinj	3/3/2020 4:14 AM
dpost	3/3/2020 4:15 AM
sinj	3/3/2020 4:14 AM

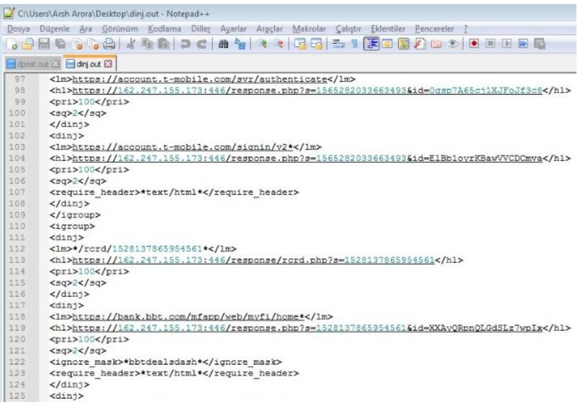
Fig.13. Encrypted dinj, dpost and sinj files



C:\Users\Arsh Arora\Desktop\dpost.out - Notepad++
Dosya Düzenle Ara Görünüm Kodlama Diller Ayarlar Araçlar Makrolar
dpost.out
dinj.out

```
QEOTNULNUL¢BSNULEOT<dpost>
<handler>http://203.176.135.102:8082</handler>
<handler>http://112.78.164.34:8082</handler>
<handler>http://103.94.122.254:8082</handler>
<handler>http://170.238.117.187:8082</handler>
<handler>http://190.100.16.210:8082</handler>
<handler>http://190.119.180.226:8082</handler>
<handler>http://96.9.77.142:80</handler>
<handler>http://96.9.73.73:80</handler>
<handler>http://36.89.106.69:80</handler>
<handler>http://177.74.232.124:80</handler>
<handler>http://103.84.238.3:80</handler>
<handler>http://194.5.250.188:443</handler>
<handler>http://217.12.209.163:443</handler>
<handler>http://185.98.87.94:443</handler>
<handler>http://92.38.135.164:443</handler>
<handler>http://212.80.218.64:443</handler>
<handler>http://64.44.51.106:443</handler>
<handler>http://146.185.253.123:443</handler>
<handler>http://185.99.2.164:443</handler>
<handler>http://51.89.115.117:443</handler>
<handler>http://188.209.52.162:443</handler>
<handler>http://185.164.32.107:443</handler>
<handler>http://107.172.191.12:443</handler>
<handler>http://51.89.115.123:443</handler>
</dpost>
›v"šDC4¯6Êi-DC43mjé°Ù¤ŠŒ''ûX—ði.-L@^™£GS]cû,+EOT
```

Fig.14. Decrypted dpost.out arff file with Python



C:\Users\Arsh Arora\Desktop\dinj.out - Notepad++
Dosya Düzenle Ara Görünüm Kodlama Diller Ayarlar Araçlar Makrolar Çalıştır Eklentiler Pencereler ?
dpost.out
dinj.out

```
<lm>https://account.t-mobile.com/svr/authenticate</lm>
<hl>https://162.247.155.173:446/response.php?s=1565282033663493&id=0qsp7A65cj1XJFoJf3c8</hl>
<pri>100</pri>
<sq>2</sq>
</dinj>
<dinj>
<lm>https://account.t-mobile.com/signin/v2*</lm>
<hl>https://162.247.155.173:446/response.php?s=1565282033663493&id=ElBb1oyrKBawVVCDCmva</hl>
<pri>100</pri>
<sq>2</sq>
<require_header>*text/html*</require_header>
</dinj>
</igroup>
<igroup>
<dinj>
<lm>*/rcrd/1528137865954561*</lm>
<hl>https://162.247.155.173:446/response/rcrd.php?s=1528137865954561</hl>
<pri>100</pri>
<sq>2</sq>
</dinj>
<dinj>
<lm>https://bank.bbt.com/mfapp/web/myfi/home*</lm>
<hl>https://162.247.155.173:446/response.php?s=1528137865954561&id=XXAyQRpnQLGdSLz7wpIx</hl>
<pri>100</pri>
<sq>2</sq>
<ignore_mask>*bbtdealsdash*</ignore_mask>
<require_header>*text/html*</require_header>
</dinj>
<dinj>
```

Fig.15. Decrypted dnj.out arff file with Python

We know that TrickBot and Emotet use a virtual network system that allows them to take over the victim's computer systems. TrickBot also used different modules to enter user credentials into any banking session. It mostly gets the downloaded modules and configuration files by running them on their servers. After running these modules, it also needs a network communication to accomplish its destructive goals. However, due to the encrypted content of the exchanged packets, their purpose is difficult to understand. Therefore, revealing the TrickBot and Emotet networking pattern will help us detect any viral infection in a system. We focus on the communication patterns between the TrickBot and Emotet servers and the compromised system.

TrickBot and Emotet network traffics were examined to determine network flow patterns. We will use machine learning techniques to detect the TrickBot and Emotet infection. Each traffic flow is defined by a set of statistical properties that can be calculated from one or more packages. Therefore, each stream will be characterized by the same set of attribute names, but different attribute values [13].

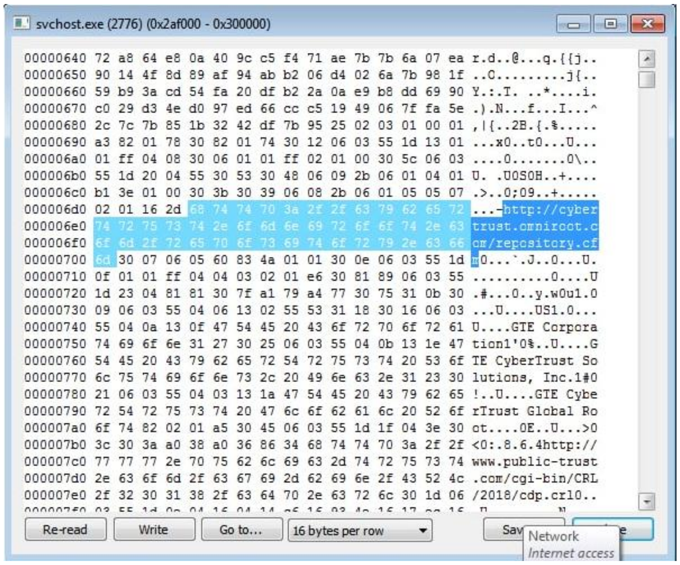
B. Machine Learning Approach to Identify TrickBot and Emotet Streams Color Space

In our methodology, we use a supervised machine learning approach to classify traffic flows by class membership. In supervised classification, classes must be predefined before the system is trained. First, a classification model is used via using a training dataset containing examples of each class. This model is then used to predict class membership for new traffic flows represented as statistical features.

We analyzed many TrickBot and Emotet malware samples statically and dynamically over 1 year period. After executing the samples of TrickBot and Emotet, a network is created between the compromised computer and a URL call is performed to reveal the public IP address of the infected computer. During the dynamic analysis, we observed the %AppData%Roaming folder to see if there were any newly created folders related to the TrickBot infection. We observed the %AppData%Local folder to see if there were any newly created folders related to the Emotet infection [14].

Before running TrickBot and Emotet malware samples in a virtual machine, we set it up to capture network traffic with some predefined filtering rules to not to capture broadcast packets with the Wireshark protocol analyzer. In Windows Task Manager it can be easily observed when the TrickBot and

Emotet process is started, finished and deployed and how many svchost processes are started by the main executable to run the downloaded module DLLs [15].

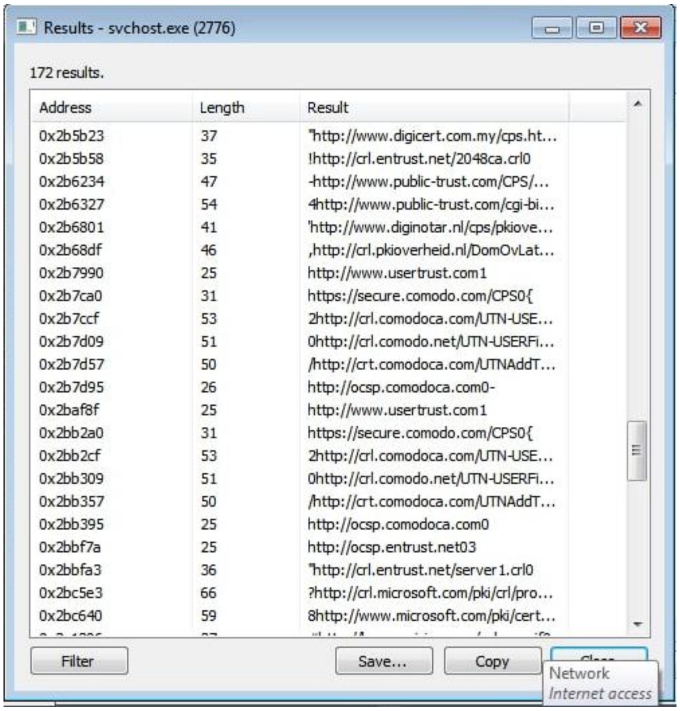


svchost.exe (2776) (0x2af000 - 0x300000)

```
00000640 72 a8 64 e8 0a 40 9c c5 f4 71 ae 7b 7b 6a 07 ea r.d..@...q.{{j..
00000650 90 14 4f 8d 89 af 94 ab b2 06 d4 02 6a 7b 98 1f ..O.........j{..
00000660 59 b9 3a cd 54 fa 20 df b2 2a 0a e9 b8 dd 69 90 Y.:.T. ..*....i.
00000670 c0 29 d3 4e d0 97 ed 66 cc c5 19 49 06 7f fa 5e .).N...f...I...^
00000680 2c 7c 7b 85 1b 32 42 df 7b 95 25 02 03 01 00 01 ,|{..2B.{.%.....
00000690 a3 82 01 78 30 82 01 74 30 12 06 03 55 1d 13 01 ...x0..t0...U...
000006a0 01 ff 04 08 30 06 01 01 ff 02 01 00 30 5c 06 03 ....0.......0\..
000006b0 55 1d 20 04 55 30 53 30 48 06 09 2b 06 01 04 01 U. .U0S0H..+....
000006c0 b1 3e 01 00 30 3b 30 39 06 08 2b 06 01 05 05 07 .>..0;09..+.....
000006d0 02 01 16 2d 68 74 74 70 3a 2f 2f 63 79 62 65 72 ...-http://cyber
000006e0 74 72 75 73 74 2e 6f 6d 6e 69 72 6f 6f 74 2e 63 trust.omniroot.c
000006f0 6f 6d 2f 72 65 70 6f 73 69 74 6f 72 79 2e 63 66 om/repository.cf
00000700 6d 30 07 06 05 60 83 4a 01 01 30 0e 06 03 55 1d m0...`.J..0...U.
00000710 0f 01 01 ff 04 04 03 02 01 e6 30 81 89 06 03 55 ..........0....U
00000720 1d 23 04 81 81 30 7f a1 79 a4 77 30 75 31 0b 30 .#...0..y.w0u1.0
00000730 09 06 03 55 04 06 13 02 55 53 31 18 30 16 06 03 ...U....US1.0...
00000740 55 04 0a 13 0f 47 54 45 20 43 6f 72 70 6f 72 61 U....GTE Corpora
00000750 74 69 6f 6e 31 27 30 25 06 03 55 04 0b 13 1e 47 tion1'0%..U....G
00000760 54 45 20 43 79 62 65 72 54 72 75 73 74 20 53 6f TE CyberTrust So
00000770 6c 75 74 69 6f 6e 73 2c 20 49 6e 63 2e 31 23 30 lutions, Inc.1#0
00000780 21 06 03 55 04 03 13 1a 47 54 45 20 43 79 62 65 !..U....GTE Cybe
00000790 72 54 72 75 73 74 20 47 6c 6f 62 61 6c 20 52 6f rTrust Global Ro
000007a0 6f 74 82 02 01 a5 30 45 06 03 55 1d 1f 04 3e 30 ot....0E..U...>0
000007b0 3c 30 3a a0 38 a0 36 86 34 68 74 74 70 3a 2f 2f <0:.8.6.4http://
000007c0 77 77 77 2e 70 75 62 6c 69 63 2d 74 72 75 73 74 www.public-trust
000007d0 2e 63 6f 6d 2f 63 67 69 2d 62 69 6e 2f 43 52 4c .com/cgi-bin/CRL
000007e0 2f 32 30 31 38 2f 63 64 70 2e 63 72 6c 30 1d 06 /2018/cdp.crl0..
```

Fig.16. ASCII codes that appear in svchost after TrickBot and Emotet viruses are executed

Svchost.exe actually stands for "service host" and it is a file used by most Windows applications. Despite this, it is generally considered a virus, as malware developers are known to add malicious files to the svchost.exe service to avoid detection. In addition, malware authors often create misspelled files such as "svhost.exe" and "svchosl.exe" to avoid detection by observers.



Results - svchost.exe (2776)

172 results.

Address	Length	Result
0x2b5b23	37	"http://www.digicert.com.my/cps.ht...
0x2b5b58	35	!http://crl.entrust.net/2048ca.crl0
0x2b6234	47	-http://www.public-trust.com/CPS/...
0x2b6327	54	4http://www.public-trust.com/cgi-bi...
0x2b6801	41	'http://www.diginotar.nl/cps/pkiove...
0x2b68df	46	,http://crl.pkioverheid.nl/DomOvLat...
0x2b7990	25	http://www.usertrust.com1
0x2b7ca0	31	https://secure.comodo.com/CPS0{
0x2b7ccf	53	2http://crl.comodoca.com/UTN-USE...
0x2b7d09	51	0http://crl.comodo.net/UTN-USERFi...
0x2b7d57	50	/http://crt.comodoca.com/UTNAddT...
0x2b7d95	26	http://ocsp.comodoca.com0-
0x2baf8f	25	http://www.usertrust.com1
0x2bb2a0	31	https://secure.comodo.com/CPS0{
0x2bb2cf	53	2http://crl.comodoca.com/UTN-USE...
0x2bb309	51	0http://crl.comodo.net/UTN-USERFi...
0x2bb357	50	/http://crt.comodoca.com/UTNAddT...
0x2bb395	25	http://ocsp.comodoca.com0
0x2bbf7a	25	http://ocsp.entrust.net03
0x2bbfa3	36	"http://crl.entrust.net/server1.crl0
0x2bc5e3	66	?http://crl.microsoft.com/pki/crl/pro...
0x2bc640	59	8http://www.microsoft.com/pki/cert...

Fig.17. Website samples obtained as a result of svchost filtering after infection

We analyzed more than 100 different TrickBot and Emotet samples over 1 year to observe how TrickBot and Emotet banking viruses evolve and discover new behavioral patterns.

With Static and Dynamic analysis, we observed the interaction of TrickBot and Emotet on our computer after they were executed We filtered the network traffic through the Wireshark protocol analyzer. We perform this because we define a network communication model and show that this interruption is caused by TrickBot and Emotet related flows in the network. During the process of capturing network traffic while the TrickBot instance is running, initial HTTP traffic is intentionally generated through interaction with popular domains. Such bona fide traffic is generated by visiting university domains, online newspaper, well-known social media websites, and some well-known websites [16].

The .pcap files captured this way also contain regular web traffic, as opposed to containing only TrickBot and Emotet-specific traffic. That's why we run each sample (infected file content) at different times. Sometimes when visiting banking web sites, it may take only a few minutes initially to observe the network traffic, and sometimes more than 2 hours to observe the injection. We captured pcap files containing both TrickBot and Emotet related traffic and also benign traffic. This difference in traffic captured in .pcap files is very useful for training and testing data for our proposed model [17]. The QUIC protocol is often observed after the Emotet virus is executed.

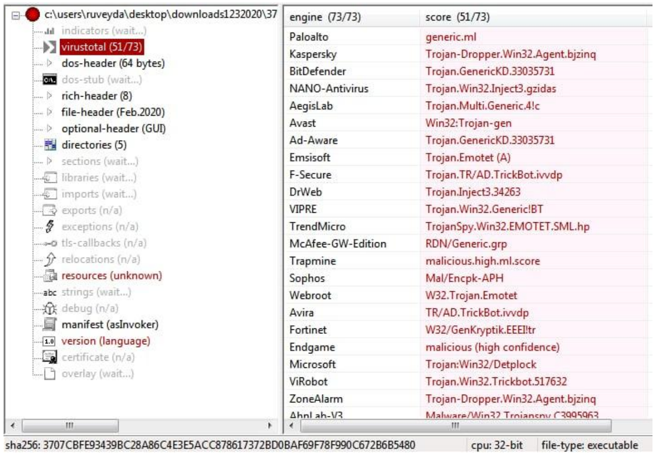
QUIC is an experimental low-latency new internet protocol implemented by Google over UDP. Generally, UDP is used in areas where speed is important and latency is not tolerated.QUIC is a protocol developed by Google. QUIC supports replicated link aggregation and aims to provide secure data transmission with similar features to TLS/SSL. It works in the same structure as the HTTP/2 protocol, but contains features that the HTTP/2 protocol cannot provide [18].

QUIC has taken a new approach to reduce latency by addressing the problems of packet loss and long RTTs (Round-round Times). It manages the former using ubiquitous TCP with UDP (User Datagram Protocol) and then minimizes the number of round trips between the sender and receiver. TCP-based delays on websites sometimes exceed milliseconds and reach up to seconds. This is where Google's new protocol QUIC comes into play. For this reason, Emotet generally prefers to exchange packets over this protocol [19].

III. IMPLEMENTATION

It is very difficult to reveal the signatures of banking viruses, which can be transmitted via e-transmitted during any banking transaction. The methodology we developed here is related to the detection of banking viruses that are harmful in such cases. While creating our classification model, we defined a data set in Excel environment to summarize different characteristics. While creating our dataset, we benefited from traffic flows and also HTTP adresses in svchost.

Our data set was created by examining the protocols. During the dataset collection phase, we collected 41437 samples from different sources including Contagio security block, MalDozer, VirusTotal, AMD datasets.



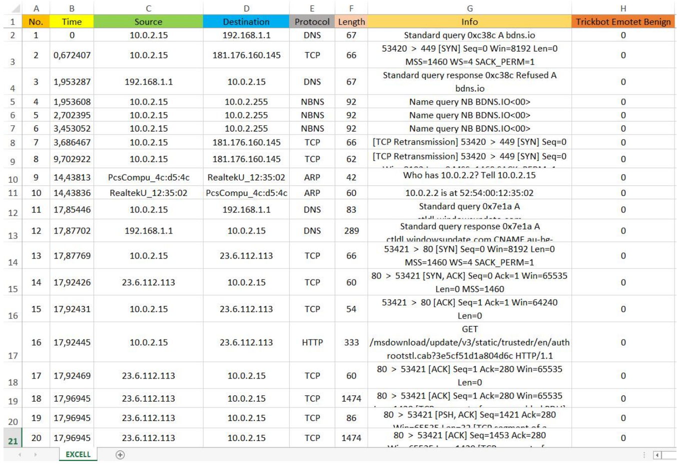
```
c:\users\ruveyda\desktop\downloads1232020\37
  indicators (wait...)
  virustotal (51/73)
  dos-header (64 bytes)
  dos-stub (wait...)
  rich-header (8)
  file-header (Feb.2020)
  optional-header (GUI)
  directories (5)
  sections (wait...)
  libraries (wait...)
  imports (wait...)
  exports (n/a)
  exceptions (n/a)
  tls-callbacks (n/a)
  relocations (n/a)
  resources (unknown)
  strings (wait...)
  debug (n/a)
  manifest (asInvoker)
  version (language)
  certificate (n/a)
  overlay (wait...)
```

engine (73/73)	score (51/73)
Paloalto	generic.ml
Kaspersky	Trojan-Dropper.Win32.Agent.bjzinq
BitDefender	Trojan.GenericKD.33035731
NANO-Antivirus	Trojan.Win32.Inject3.gzidas
AegisLab	Trojan.Multi.Generic.4!c
Avast	Win32:Trojan-gen
Ad-Aware	Trojan.GenericKD.33035731
Emsisoft	Trojan.Emotet (A)
F-Secure	Trojan.TR/AD.TrickBot.ivvdp
DrWeb	Trojan.Inject3.34263
VIPRE	Trojan.Win32.Generic!BT
TrendMicro	TrojanSpy.Win32.EMOTET.SML.hp
McAfee-GW-Edition	RDN/Generic.grp
Trapmine	malicious.high.ml.score
Sophos	Mal/Encpk-APH
Webroot	W32.Trojan.Emotet
Avira	TR/AD.TrickBot.ivvdp
Fortinet	W32/GenKryptik.EEEI!tr
Endgame	malicious (high confidence)
Microsoft	Trojan:Win32/Detplock
ViRobot	Trojan.Win32.Trickbot.517632
ZoneAlarm	Trojan-Dropper.Win32.Agent.bjzinq
AhnLab-V3	Malware/Win32.Trojanspy.C3995963

sha256: 3707CBFE93439BC28A86C4E3E5ACC878617372BD0BAF69F78F990C672B6B5480 | cpu: 32-bit | file-type: executable

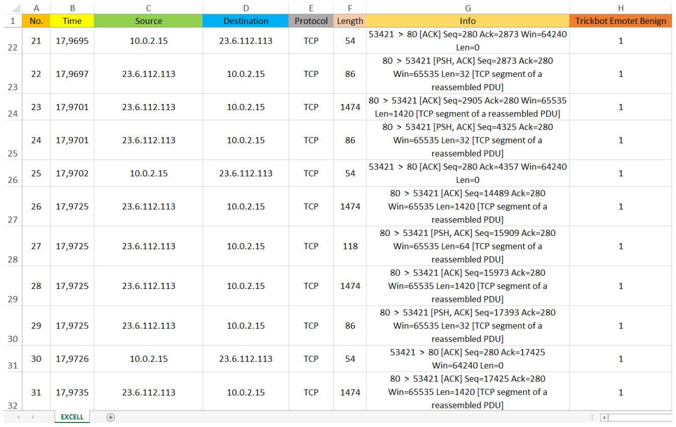
Fig.18. Infected IPs discovered as a result of Vırustotal analysis

Due to the analysis performed, 13077 out of 41437 samples were determined as working samples. After data collection from December 2021 to December 2022, we categorized the data according to its functionality whether it was malware and, if so, what type. As a result, we obtained a set containing 13077 data, consisting of a total of 5 categories. Our dataset has 9803 row, including Benign flows, Banking Malware (Emotet, TrickBot). The row number of benign flows are 1795 and it is marked as in the benign software category. They also presented as two different sets which contain 470 and 139 features. In this study, datasets containing 500 extracted features were used. Below, we display the data set, which contain 10 Benign, 10 TrickBot and 10 Emotet samples.



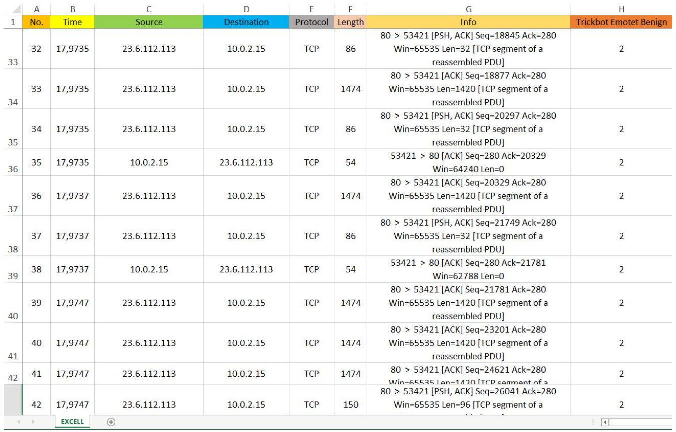
No.	Time	Source	Destination	Protocol	Length	Info	Trickbot Emotet Benign
1	0	10.0.2.15	192.168.1.1	DNS	67	Standard query 0xc38c A bdns.io	0
2	0,672407	10.0.2.15	181.176.160.145	TCP	66	53420 > 449 [SYN] Seq=0 Win=8192 Len=0 MSS=1460 WS=4 SACK_PERM=1	0
3	1,953287	192.168.1.1	10.0.2.15	DNS	67	Standard query response 0xc38c Refused A bdns.io	0
4	1,953608	10.0.2.15	10.0.2.255	NBNS	92	Name query NB BDNS.IO<00>	0
5	2,702395	10.0.2.15	10.0.2.255	NBNS	92	Name query NB BDNS.IO<00>	0
6	3,453052	10.0.2.15	10.0.2.255	NBNS	92	Name query NB BDNS.IO<00>	0
7	3,686467	10.0.2.15	181.176.160.145	TCP	66	[TCP Retransmission] 53420 > 449 [SYN] Seq=0	0
8	9,702922	10.0.2.15	181.176.160.145	TCP	62	[TCP Retransmission] 53420 > 449 [SYN] Seq=0	0
9	14,43813	PcsCompu_4c:d5:4c	RealtekU_12:35:02	ARP	42	Who has 10.0.2.2? Tell 10.0.2.15	0
10	14,43836	RealtekU_12:35:02	PcsCompu_4c:d5:4c	ARP	60	10.0.2.2 is at 52:54:00:12:35:02	0
11	17,85446	10.0.2.15	192.168.1.1	DNS	83	Standard query 0x7e1a A	0
12	17,87702	192.168.1.1	10.0.2.15	DNS	289	Standard query response 0x7e1a A ctldl windowsupdate com CNAME au-bg-	0
13	17,87769	10.0.2.15	23.6.112.113	TCP	66	53421 > 80 [SYN] Seq=0 Win=8192 Len=0 MSS=1460 WS=4 SACK_PERM=1	0
14	17,92426	23.6.112.113	10.0.2.15	TCP	60	80 > 53421 [SYN, ACK] Seq=0 Ack=1 Win=65535 Len=0 MSS=1460	0
15	17,92431	10.0.2.15	23.6.112.113	TCP	54	53421 > 80 [ACK] Seq=1 Ack=1 Win=64240 Len=0	0
16	17,92445	10.0.2.15	23.6.112.113	HTTP	333	GET /msdownload/update/v3/static/trustedr/en/auth rootstl.cab?3e5cf51d1a804d6c HTTP/1.1	0
17	17,92469	23.6.112.113	10.0.2.15	TCP	60	80 > 53421 [ACK] Seq=1 Ack=280 Win=65535 Len=0	0
18	17,96945	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=1 Ack=280 Win=65535	0
19	17,96945	23.6.112.113	10.0.2.15	TCP	86	80 > 53421 [PSH, ACK] Seq=1421 Ack=280	0
20	17,96945	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=1453 Ack=280	0

Fig.19.a.Dataset example Benign flows (except TrickBot and Emotet)



No.	Time	Source	Destination	Protocol	Length	Info	Trickbot Emotet Benign
21	17,9695	10.0.2.15	23.6.112.113	TCP	54	53421 > 80 [ACK] Seq=280 Ack=2873 Win=64240 Len=0	1
22	17,9697	23.6.112.113	10.0.2.15	TCP	86	80 > 53421 [PSH, ACK] Seq=2873 Ack=280 Win=65535 Len=32 [TCP segment of a reassembled PDU]	1
23	17,9701	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=2905 Ack=280 Win=65535 Len=1420 [TCP segment of a reassembled PDU]	1
24	17,9701	23.6.112.113	10.0.2.15	TCP	86	80 > 53421 [PSH, ACK] Seq=4325 Ack=280 Win=65535 Len=32 [TCP segment of a reassembled PDU]	1
25	17,9702	10.0.2.15	23.6.112.113	TCP	54	53421 > 80 [ACK] Seq=280 Ack=4357 Win=64240 Len=0	1
26	17,9725	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=14489 Ack=280 Win=65535 Len=1420 [TCP segment of a reassembled PDU]	1
27	17,9725	23.6.112.113	10.0.2.15	TCP	118	80 > 53421 [PSH, ACK] Seq=15909 Ack=280 Win=65535 Len=64 [TCP segment of a reassembled PDU]	1
28	17,9725	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=15973 Ack=280 Win=65535 Len=1420 [TCP segment of a reassembled PDU]	1
29	17,9725	23.6.112.113	10.0.2.15	TCP	86	80 > 53421 [PSH, ACK] Seq=17393 Ack=280 Win=65535 Len=32 [TCP segment of a reassembled PDU]	1
30	17,9726	10.0.2.15	23.6.112.113	TCP	54	53421 > 80 [ACK] Seq=280 Ack=17425 Win=64240 Len=0	1
31	17,9735	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=17425 Ack=280 Win=65535 Len=1420 [TCP segment of a reassembled PDU]	1

Fig.19.b.Dataset example for TrickBot



No.	Time	Source	Destination	Protocol	Length	Info	Trickbot Emotet Benign
32	17,9735	23.6.112.113	10.0.2.15	TCP	86	80 > 53421 [PSH, ACK] Seq=18845 Ack=280 Win=65535 Len=32 [TCP segment of a reassembled PDU]	2
33	17,9735	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=18877 Ack=280 Win=65535 Len=1420 [TCP segment of a reassembled PDU]	2
34	17,9735	23.6.112.113	10.0.2.15	TCP	86	80 > 53421 [PSH, ACK] Seq=20297 Ack=280 Win=65535 Len=32 [TCP segment of a reassembled PDU]	2
35	17,9735	10.0.2.15	23.6.112.113	TCP	54	53421 > 80 [ACK] Seq=280 Ack=20329 Win=64240 Len=0	2
36	17,9737	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=20329 Ack=280 Win=65535 Len=1420 [TCP segment of a reassembled PDU]	2
37	17,9737	23.6.112.113	10.0.2.15	TCP	86	80 > 53421 [PSH, ACK] Seq=21749 Ack=280 Win=65535 Len=32 [TCP segment of a reassembled PDU]	2
38	17,9737	10.0.2.15	23.6.112.113	TCP	54	53421 > 80 [ACK] Seq=280 Ack=21781 Win=62788 Len=0	2
39	17,9747	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=21781 Ack=280 Win=65535 Len=1420 [TCP segment of a reassembled PDU]	2
40	17,9747	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=23201 Ack=280 Win=65535 Len=1420 [TCP segment of a reassembled PDU]	2
41	17,9747	23.6.112.113	10.0.2.15	TCP	1474	80 > 53421 [ACK] Seq=24621 Ack=280 Win=65535 Len=1420 [TCP segment of a	2
42	17,9747	23.6.112.113	10.0.2.15	TCP	150	80 > 53421 [PSH, ACK] Seq=26041 Ack=280 Win=65535 Len=96 [TCP segment of a	2

Fig.19.c.Dataset example for Emotet

The analysis of the dataset using machine-learning classifiers was carried out with the WEKA software which was developed at the University of Waikato. It is the abbreviation for Waikato Environment for Knowledge Analysis. This code, which is a JAVA open source library, contains an algorithm that can be applied to devices with Android operating system [20]. In the classification results made with WEKA, False Positive Ratio (FPR), True Positive Ratio (TPR), Precision, Recall, FMeasure etc. values are given. These values are an important criterion in interpreting the results. TPR, correctly defined data; FPR, misidentified data; Precision is expressed as the ratio of the correct data of a category to the incorrect data of that category and is formulated as follows [20, 21].

$$TPR = TP\ FN{+}TP \tag{1}$$

$$FPR = FP\ TN{+}FP \tag{2}$$

$$Precision = TP\ TP{+}FP \tag{3}$$

$$F - Measure = 2{*}TP\ 2{*}TP{+}FP{+}FN \tag{4}$$

$$Recall{=}\ TP\ FN{+}TP \tag{5}$$

$$Accuracy(\%) = TN{+}TP\ TP{+}FN{+}FP{+}TN * 100 \tag{6}$$

The path followed in the study is given in Figure 20. Up to this point, the dataset and WEKA evaluation criterias are included. The next steps are covered in the 'Results and Discussion' section in detail.

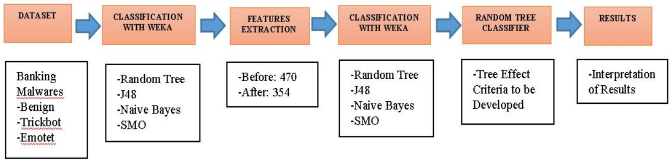
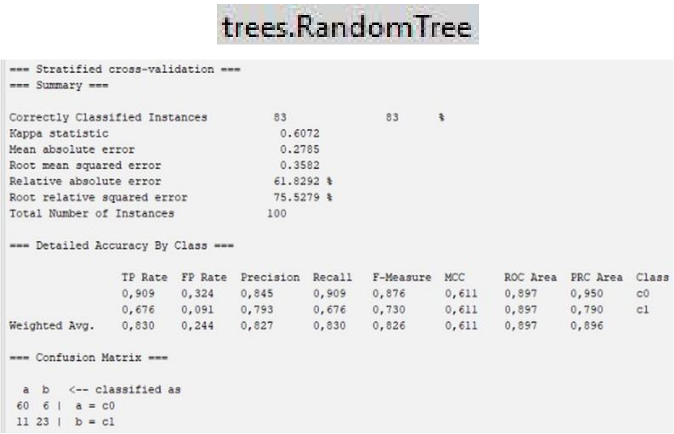


Fig.20. Flow chart of study

In this study, in which analyzes were made for the detection of malware, first of all, the data set was first analyzed using machine learning (ML) classifiers such as SMO, Naive Bayes (NB), J48 and Random Tree (RT) algorithms. Then, feature extraction was performed and the results were compared with the same ML classifiers. The effect of different parameters was examined by using the algorithm that gave the best results.

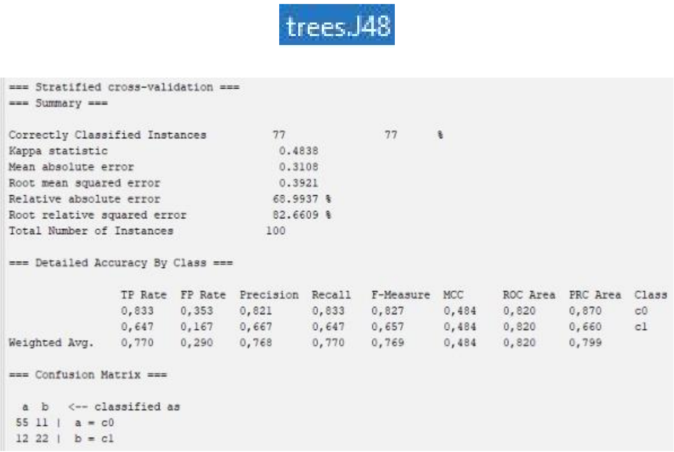
A. *Effects of Algorithms*

When the literature is examined, it is seen that ML (Machine Learning) algorithms are frequently used in malware detection. In this context, 4 different classifiers, namely SMO, NB (Naive Bayes), J48 and RT (Random Tree), were included in the study. Classification results using the WEKA program are given in Figures 21, 22, 23 and 24. It is seen that the Random Tree classifier is the algorithm that gives the best result with a success rate of 83% in the success evaluation using the accuracy percentage. The lowest success is the SMO algorithm with 60%. Results for J48 and NB were 77% and 64%, respectively. TPR, FPR etc. given in WEKA analysis outputs in Figure 25. The results of the criteria are given. All classifiers, SMS Malware appear to have high accuracy. This result is consistent with the findings from the study [22].


trees.RandomTree

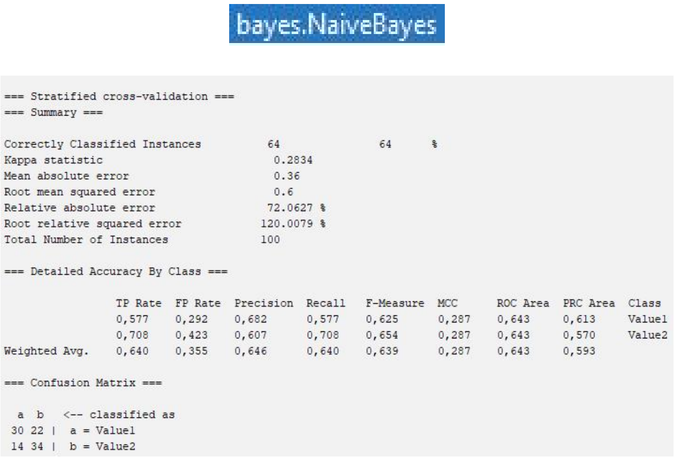
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          83               83      %
Kappa statistic                          0.6072
Mean absolute error                      0.2785
Root mean squared error                  0.3582
Relative absolute error                 61.8292 %
Root relative squared error             75.5279 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,909    0,324    0,845      0,909    0,876      0,611    0,897     0,950     c0
                 0,676    0,091    0,793      0,676    0,730      0,611    0,897     0,790     c1
Weighted Avg.    0,830    0,244    0,827      0,830    0,826      0,611    0,897     0,896

=== Confusion Matrix ===

  a  b   <-- classified as
 60  6 |  a = c0
 11 23 |  b = c1
```

Fig.21. Random Tree classification algorithm example result in WEKA application


trees.J48

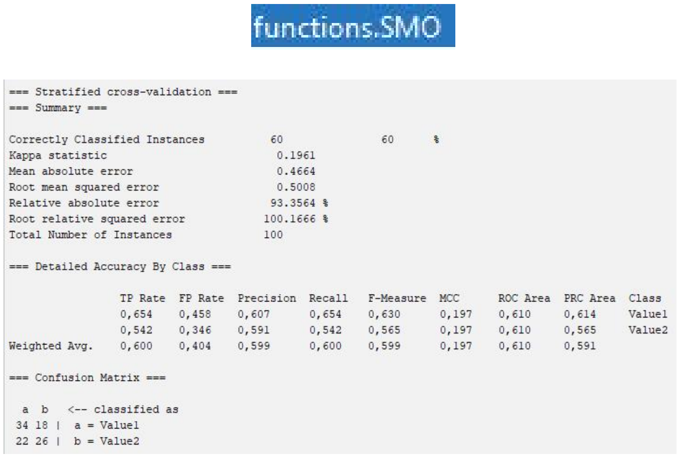
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          77               77      %
Kappa statistic                          0.4838
Mean absolute error                      0.3108
Root mean squared error                  0.3921
Relative absolute error                 68.9937 %
Root relative squared error             82.6609 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,833    0,353    0,821      0,833    0,827      0,484    0,820     0,870     c0
                 0,647    0,167    0,667      0,647    0,657      0,484    0,820     0,660     c1
Weighted Avg.    0,770    0,290    0,768      0,770    0,769      0,484    0,820     0,799

=== Confusion Matrix ===

  a  b   <-- classified as
 55 11 |  a = c0
 12 22 |  b = c1
```

Fig.22. Trees.J48 classification algorithm example result in WEKA application


bayes.NaiveBayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          64               64      %
Kappa statistic                          0.2834
Mean absolute error                      0.36
Root mean squared error                  0.6
Relative absolute error                 72.0627 %
Root relative squared error            120.0079 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,577    0,292    0,682      0,577    0,625      0,287    0,643     0,613     Value1
                 0,708    0,423    0,607      0,708    0,654      0,287    0,643     0,570     Value2
Weighted Avg.    0,640    0,355    0,646      0,640    0,639      0,287    0,643     0,593

=== Confusion Matrix ===

  a  b   <-- classified as
 30 22 |  a = Value1
 14 34 |  b = Value2
```

Fig.23. Naive Bayes classification algorithm example result in WEKA application


functions.SMO

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          60               60      %
Kappa statistic                          0.1961
Mean absolute error                      0.4664
Root mean squared error                  0.5008
Relative absolute error                 93.3564 %
Root relative squared error            100.1666 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,654    0,458    0,607      0,654    0,630      0,197    0,610     0,614     Value1
                 0,542    0,346    0,591      0,542    0,565      0,197    0,610     0,565     Value2
Weighted Avg.    0,600    0,404    0,599      0,600    0,599      0,197    0,610     0,591

=== Confusion Matrix ===

  a  b   <-- classified as
 34 18 |  a = Value1
 22 26 |  b = Value2
```

Fig.24. SMO classification algorithm example result in WEKA application

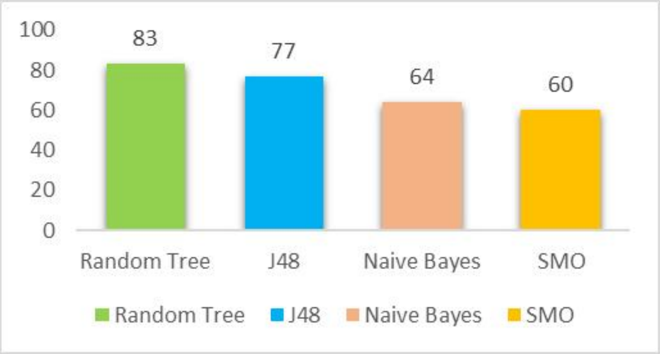
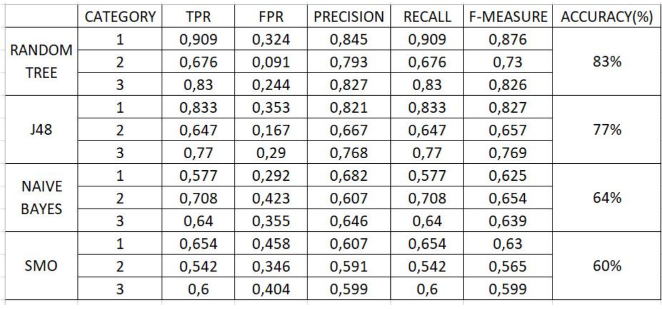


Fig.25. Classifier accuracy



	CATEGORY	TPR	FPR	PRECISION	RECALL	F-MEASURE	ACCURACY(%)
RANDOM TREE	1	0,909	0,324	0,845	0,909	0,876	83%
	2	0,676	0,091	0,793	0,676	0,73	
	3	0,83	0,244	0,827	0,83	0,826	
J48	1	0,833	0,353	0,821	0,833	0,827	77%
	2	0,647	0,167	0,667	0,647	0,657	
	3	0,77	0,29	0,768	0,77	0,769	
NAIVE BAYES	1	0,577	0,292	0,682	0,577	0,625	64%
	2	0,708	0,423	0,607	0,708	0,654	
	3	0,64	0,355	0,646	0,64	0,639	
SMO	1	0,654	0,458	0,607	0,654	0,63	60%
	2	0,542	0,346	0,591	0,542	0,565	
	3	0,6	0,404	0,599	0,6	0,599	

Fig.26. Classification optimization results from the WEKA

In Figure.26, the data numbers classified according to the categories are given. Category 1 (Benign), Category 2 (TrickBot), Category 3 (Emotet) are expressed as 1, 2, 3, respectively. Highest result (Random Tree- 83% vs. J48 - 77%), compared to the number of benign software detected in J48 (1459) is higher than that found in Random Tree (1379). Sum of numbers for each category data is 10607 for a correctly classified Random Tree. In J48, the correct classification result in all categories is as follows: 10181. In short, although the success of detecting benign software in J48 was 77% (83% for Random Tree), considering the overall percentage of accuracy, Random Tree appears to be a better classifier under these categories.

B. Feature Extraction Effect

Feature reduction is one of the key pieces of work in malware detection. In this study, 116 features with the lowest effect on the ranking were removed and reconstructed. We count 470 feature for the analysis. The results obtained for the RT, NB, J48 and SMO classifiers, compared with the results before feature extraction (Figure 27).

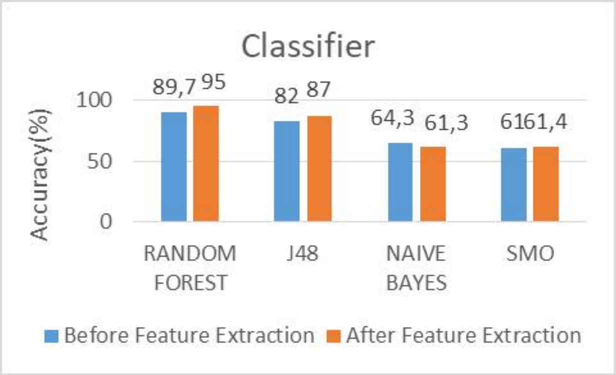


Fig.27. Accuracy comparison before and after feature extraction.

In Figure 27, the change in accuracy for the Random Tree classifier after feature extraction was minimal. The greatest increase in the accuracy of the results was observed for NB. Contrary to the others, there is a small decrease in J48. In the analysis made so far, the success of different classifiers and the effect of feature extraction in malware detection have been examined. In the comparison, as seen in Figure 25, the malware was tagged with the best Random Tree classifier. Based on this achievement, analyzes were made for the Random Tree classifier and 354 features in the next parts of the study.

C. Tree Effect Criteria to be Developed

According to the findings of the study, Random Tree is the best performing classifier for detecting banking malicious software (TrickBot and Emotet). Among other classifiers, new analyzes were made by changing the number based on this information.

Random forest algorithm in Random Tree algorithm is one of the supervised classification algorithms. It is used in both regression and classification problems. The algorithm aims to increase the classification value during the classification process by producing more than one decision tree. Random forest algorithm is the process of choosing the highest score among many decision trees that work independently of each other. As the number of trees (our data) increases, our rate of obtaining a precise result increases [23]. The main difference between the decision tree algorithm and the random forest algorithm is that the process of finding the root node and splitting the nodes is random. The random forest algorithm reduces the problem of over-learning if you have enough trees. It requires little data preparation. It requires little data preparation. It requires little data preparation. The aim is to observe ın the algorithm with the best classification, different parameters are tried to determine the accuracy and reach the final result.

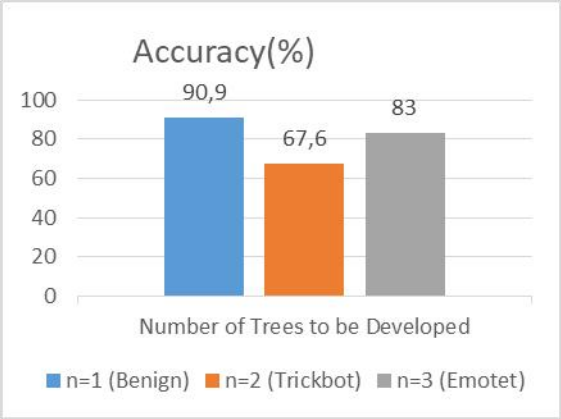


Fig.28. Tree Effect Criteria to be Developed

Figure 28 examines the effect of the number of trees to be developed for the Random Tree classifier. For Benign, Emotet and TrickBot an increase in n indicates the impact on malware detection.

It is concluded that for malware detection, the Random Tree classifier determines the best discrimination.

IV. Conclusion

The information age that we live in has brought along some problems as well as providing great convenience for humanity. As the access to information, technology and internet became easier, malicious use of internet also has become a significant problem. In parallel with the increase in these threats, which pose a great danger to information security, prevention and detection activities in these areas have also accelerated. In the field of information security, malware detection studies, which are also frequently encountered in the academic world, are conducted to identify threats developed with malicious intent. In this study, we develop a technique by using ML classifiers to determine the banking trojan infection.

The percentage of success in malware detection studies is associated with accurate detection of malwares. Tagging good software as malicious software can cost money and time. However, since labeling malware as benign will cause even greater damage. In this context, it has been seen that tagging malware correctly has improved the reliability of our study. Also in addition to the effect of feature extraction, we also study the classifier performance. According to the findings of these two phases, the best classification success (before and after feature extraction) belongs to Random Tree algorithm for TrickBot and Emotet detection. The change in the number of trees has provided the desired success in malware detection.

In our analysis, we observe that Random Tree and J48 give better results compared to other detection techniques. Despite higher flow detection with J48, Random Tree performed better overall. We obtained 83% Our dataset, which we ran in the Weka program, yielded the following results: Random Tree 83%, J48 77%, Naive Bayes 64% and SMO 60%.

In short, in this age where information is under threat, malware detection and prevention is of great importance. Detection of banking malicious software is one of the shining areas for the security of banking customers. This study has enriched the literature in terms of examining this correct labeling. After evaluating the classifier performance and feature extraction efficiency, Random Tree gives best results in terms of classification of benign TrickBot and Emotet traffic flows.

References

[1] M. Edwin Agwu, “Analysis of Obstacles to Uptake of Internet Banking Services in Nigeria” Research Journal & Management-RJBM (2015), Vol.2(1)doi:10.17261/Pressacademia.201519824 Available: https://dergipark.org.tr/tr/download/article-file/375170

[2] M. Zainab Alkhalil, Chaminda Hewage “Phishing Attacks: A Recent Comprehensive Study and a New Anatomy” Liqaa Nawaf and Imtiaz Khan Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, United Kingdom Front. Comput. Sci., 09 March 2021 Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.563060/full

[3] Debbie Walkowski “Banking Trojans: A Reference Guide to the Malware Family Tree By Remi Cohen Additional Contributions” August 09, 2019 Available: https://www.f5.com/labs/articles/education/banking-trojans-a-reference-guide-to-the-malware-family-tree

[4] Cybersecurity and Infrastructure Security Team “Emotet Malware” July 20, 2018 Last Revised: January 23, 2020 Available: https://us-cert.cisa.gov/ncas/alerts/TA18-201A

[5] Michelle Drolet “What is Emotet? And how to guard against this Persistent Trojan Malware” Contributor, April 12, 2019 Available: https://www.csoonline.com/article/3387146/what-is-emotet-and-how-to-guard-against-this-persistent-trojan-malware.html

[6] R. Çelik, A. Gezer “Behavioral Analysis of Tricot Banking Trojan with its New Tricks” International Journal of Technology and Engineering Studies.Available:https://kkgpublications.com/wpcontent/uploads/2019/12/ijtes.5.10004-3.pdf

[7] Alexander S. Gillis, K. Elissa “TrickBot Malware, After Emotet takedown, TrickBot roars up threat charts” Technical Writer and Editor in ComputerWeekly Available:https://www.techtarget.com/searchsecurity/definition/TrickBot-malware

[8] David Garcia “Vadokrist: Banking Malware Targeting Brazilian Entities” Fer. 17, 2020 Available:https://www.revelock.com/en/blog/vadokrist-banking-malware-targeting-brazilian-entities

[9] Revti Vadjikar “Top 4 Ways Emotet Breaches Banking Security” Factspan, January 15, 2018 Available: https://www.factspan.com/top-4-ways-emotet-breaches-banking-security

[10] PCrisk Team “Emotet Blunders through Attack Campaign” PCrisk, 24 September 2020 Available: https://www.pcrisk.com/internet-threat-news/18929-emotet-blunders-through-attack-campaign

[11] A Gezer, G Warner, C Wilson, P Shrestha “A flow-based approach for TrickBot banking trojan detection” Computers & Security, 2019 Elsevier Van Bladel, *Electromagnetic Fields*, John Wiley & Sons, 2007, p.1176. Available:https://aperta.ulakbim.gov.tr/record/111954#.YZXwNWBByUk

[12] Aditya K. Sood, Richard Enbody “Multi-staged Attacks Driven by Exploits and Malware” in Targeted Cyber Attacks Malware Infection, April18,2014 Available:https://www.elsevier.com/books/targeted-cyber-attacks/sood/978-0-12-800604-7

[13] Malware Analysis by Hasherezade on 29 Dec 2017

Available:https://github.com/hasherezade/malware_analysis/tree/master/trickbot

[14] Frederick Lardinois "Google Wants to Speed Up the Web With Its Quic Protocol" Techcrunch, 3 April 18, 2015 Available: https://www.ajer.org/papers/v6(04)/F06044045.pdf

[15] Steve Patrick "Network protocols, What are QUIC? Everything You Need to Know" in APNIC, September 14, 2021 Available: https://blog.apnic.net/2019/03/04/a-quick-look-at-quic/

[16] Leslie F. Sikos "Forensic Science International: Digital Investigation" Volume 32, March 2020, 200892 Available:https://www.sciencedirect.com/journal/forensic-science-international-digital-investigation/vol/32/suppl/C

[17] Sunghoon Lee "Using Weka in Matlab" version 1.5 Mathworks Jul 22, 2015 Available:https://www.mathworks.com/matlabcentral/fileexchange/50120-using-weka-in-matlab

[18] Nir Shwarts, Kessem L. "Trojan Widens Its Attack Scope in Spain, Brings Redirection Attacks to Local Banks" Security Intelligence, July 19, 2017 Available:https://www.imperva.com/learn/application-security/dns-hijacking-redirection/

[19] Katsumi Ono, Isamu Kawaishi, Toshihiko Kamon "Trend of Botnet Activities" Proceedings of the 41st Annual IEEE International Carnahan Conference on Security Technology, Canada (2007), pp. 243-249, November 2007 Available: https://ieeexplore.ieee.org/document/4373496

[20] J. Davison "TrickBot Banking Trojan Adapts With The New Module" Webroot Threat Lab, March 21, 2018 Available: https://www.webroot.com/blog/2018/03/21/trickbot-banking-trojan-adapts-new-module/

[21] Marc Salinas, Jose Miguel Holguin "Innovation in Process Malware Report Evolution of TrickBot" June, 2017 Available:https://www.slideshare.net/rootedcon/jose-miguel-holguin-marc-salinas-taller-de-anlisis-de-memoria-ram-en-sistemas-windows-rooted2019

[22] Liu, J., Xiao, Y., Ghaboosi, K., Deng, H., Zhang, J. "Botnet: Classification, Attacks, Detection, Tracing, and Preventive Measures" EURASIP Journal on Wireless Communications and Networking. Volume 2009. Available: https://doi.org/10.1155/2009/69265

[23] Yusuf Sönmez, Meltem Salman and Murat Dener "Performance Analysis of Machine Learning Algorithms for Malware Detection by Using CICMalDroid2020 Dataset" Available:https://dergipark.org.tr/tr/download/article-file/2060165

BIOGRAPHIES

RÜVEYDA ÇELİK was born in Kayseri City, Turkey, in 1996. Sreceived bachelor degree in Electrical and Electronics Engineering from Niğde Ömer Halisdemir University Niğde, TURKEY, in 2019. She started master's degree in Electrical and Electronics Engineering at Kayseri University in 2019 and still continues.

Her research interests include internet traffic analysis, network traffic modelling and characterization, malicious banking viruses, data mining, telecommunication technologies, IoT systems, and malware analysis.

ALİ GEZER was born in Kayseri City, Turkey, in 1976. He received the B.S. degree in Electronic and Computer Education from Marmara University in 1999 and M.S. degree in Computer Engineering from Erciyes University in 2004, and the Ph.D. degree in Electronic Engineering from Erciyes University, Kayseri, TURKEY, in 2011. He is an Associated Professor with the Electronic and Communication Technology in Kayseri University.

His research interests include internet traffic analysis, self-similarity, network traffic modelling and characterization, signal processing techniques, telecommunication technologies, IoT botnet investigations, and malware analysis.

A Novel Approach to Improve the Performance of the Database Storing Big Data with Time Information

Murat Tasyurek

Abstract— **Big data is defined as data sets that are too large and/or complex to be processed by classical data processing methods. Big data analysis is essential because it enables more competent business movements, more efficient operations, and higher profits by using the data of institutions and organizations. However, large data sets are difficult to analyze because they are produced quickly, require large storage areas in computer systems, and the diversity of their data. In this study, a new approach using the denormalization method is proposed to accelerate the response time of the database in database systems where large volumes of data containing historical information are stored. Denormalization is defined as the process of adding rows or columns that are not needed to increase the reading performance of the database to the database system that has been normalized. In the proposed approach in this study, a large-volume data set consisting of real spatial data belonging to Kayseri Metropolitan Municipality (KBB), containing temporal information and having approximately 96,000,000 row records, was used. In the proposed approach, the response time of the query is accelerated by recording the time information as numbers to increase the query performance of large volumes of data recorded in date format due to the temporal query process. The performance of the proposed method is compared with the performance of the normalization method using actual data on Microsoft SQL Server and Oracle database systems. The method proposed in the experimental evaluations shows that it works approximately eight times faster. In addition, the experimental results showed that the proposed method improves query performance more than the normalization-based method as the data size increases.**

Index Terms— **Database performance, Denormalization, Large volume data, Temporal query.**

I. Introduction

WITH THE widespread use of technology, data with different attributes are formed in many application areas [1]. For instance, data sent by smartphones and sensors, surveys, publicly shared comments (images, videos) on websites are data with different attributes [2]. If data is produced quickly, needs large storage areas, or consists of data with various features, it shows big data characteristics [3]. Big data is defined as data sets that are too large and/or complex to be processed by classical data processing methods [4]. Big data is usually stored in databases and analyzed using software specifically designed to handle large and complex data sets [5]. Big data analysis is vital because it enables more competent business movements, more efficient operations, and higher profits by using the data of institutions and organizations [6]. However, large data sets are difficult to analyze because they are produced quickly, require large storage areas in computer systems, and the diversity of their data [7].

The data production rate, size (volume), and diversity of big data do not mean anything on their own [8]. For the data set to be useful or usable, it must be transformed into information using various techniques. To analyze the data, it must be saved in a file or database system in a specific format. Database systems are widely used because they are designed to easily organize, store and retrieve large amounts of data [9]. The normalization method known as relational database theory is used in database design [10]. Normalization, also known as parsing, parses a table consisting of many columns and rows into subsets with fewer rows and columns to avoid duplication [11]. However, the increase in the number of records and tables in the database increases the response time of the database. To accelerate the response time of the database system to the queries, the process of accelerating the response time of the database by adding rows or columns that are not needed after the normalization process in the database design is defined as denormalization [12, 13]. In this study, a new denormalization-based method was proposed to accelerate the analysis of the data stored in the relational database, which consists of actual spatial data belonging to the KBB, contains time information, approximately 96,000,000 row records. In this study, the performance of the proposed novel approach is compared with the performance of the normalization-based method by storing the actual data in Oracle and MS SQL database systems.

A. Contributions

- In this study, a new approach is proposed to improve the performance of database querying in which large

MURAT TAŞYÜREK, is with Department of Computer Engineering of Kayseri University, Kayseri, Turkey, (e-mail: m_tasyurek@hotmail.com).

https://orcid.org/0000-0001-5623-8577

Manuscript received Jan 17, 2022; accepted Aug 16, 2022.
DOI: 10.17694/bajece.1059070

volumes of data containing time information are stored.

- The performance of the proposed approach is compared on MS SQL and Oracle database systems using actual data from KBB.

B. Scope and Outline

- The main purpose of this study is to develop a new query technique for large databases containing space and time information with fast growth and large volume. For this reason, examining other features such as the variety that make up big data is out of the scope of this study.

In the following parts of this study, a literature review is presented, problems and approaches are introduced, the results of the experiments are discussed, and the results are shared.

II. Related Works

A database is an organized system of structured information and/or data stored digitally (electronically) in computer systems [14]. Database systems are widely used because they make it easy to access, manage, modify, update, control, and organize data [15]. To use these features of the database system and to work quickly, table designs must be created in a certain order. To ensure data integrity in the database system, dividing the data into sub-tables and to save them relationally is defined as normalization. However, as the number of tables and data increases in database systems, the response times of the queries used to extract information from the database get longer. To speed up the search process in database systems, indexing the searched columns using one of the indexing methods provided by the relevant database system is a widely used technique [16, 17]. However, since the indexing process requires a lot of storage space in computer systems, it cannot be used in cases where the storage space is limited. On the other hand, Online Analytical Processing (OLAP) method is an additional software technology that provides information by combining, grouping, or combining the attributes in databases and is used for rapid extraction of information from database systems [18, 19]. The denormalization method, known as accelerating the response time of the database by adding rows or columns that are not needed in the database system, is used instead of the OLAP system, which requires additional costs [20]. The denormalization method is a widely used technique to improve the performance of the database system [21-24]. On the other hand, NoSQL database systems, which have flexible schemas for modern applications and are specially designed for certain data models, have become widespread due to their practical use, ease of development, and performance [25, 26]. NoSQL databases use different data models, including graph, key-value, document, in-memory search [27]. On the other hand, NoSQL databases are designed for various data access patterns involving low latency applications [28, 29]. For this reason, NoSQL database systems are not preferred in cases where low latency is important.

In addition, many devices that make up the transportation system are constantly generating data. For this reason, the data obtained from the devices used in the transportation system is considered big data. Special solutions need to be developed to overcome the problems running in the transport system. For example, Vela et al. [30] focused on the problem of designing the NoSQL document database with which to manage the information concerning accessible routes obtained by means of crowdsourcing techniques. Asaithambi et al. [31] proposed the microservice oriented big data architecture incorporating data processing techniques, like predictive modelling for achieving smart transportation and analytics microservices required towards smart cities. Gonzalez et al. [32 investigated into the testing of transactional services in NoSQL databases in order to test and analyses the data consistency by taking into account the characteristics of NoSQL databases for efficiency and velocity.

In this study, a new approach is proposed to overcome the problem of low latency in NoSQL database systems and to additional costs of OLAP systems, and to improve the date-based filtering and sorting performance of large volumes of spatial data containing time information.

III. Approaches

A. Basic Problem

The problem arose in keeping and analyzing the location information of the movements of the vehicles operating in the KBB public transportation system. In the KBB public transportation fare collection system, payments are made by considering the number of kilometers traveled by the vehicles along the route. However, traffic density, etc. Due to the reasons, the start and end times of the flights do not fully comply with the planned departure times. On the other hand, some of the delays during the voyage hours are caused by errors or omissions caused by the driver. Whatever the reason, it is vital to detect and solve the problem since the delays in the public transportation system affect the citizens using this system. To analyze a voyage in the public transport system, the locations sent by the vehicle before the start time of the journey, the locations sent during the trip, and the location information sent after the end time of the voyage can be determined together. To carry out this analysis, the location information (latitude and longitude), line information, speed, vehicle side number, and time information sent by the vehicles working in the public transportation system were recorded in the database system used by the KBB using the normalization method. In Fig. 1, the voyage start time, voyage duration, number of passengers carried, end time, and location information sent along the route of a vehicle with side number 387 operating on route 563 belonging to the KBB transportation system are presented on the map. The icons shown in Fig. 1 represent the following meanings:

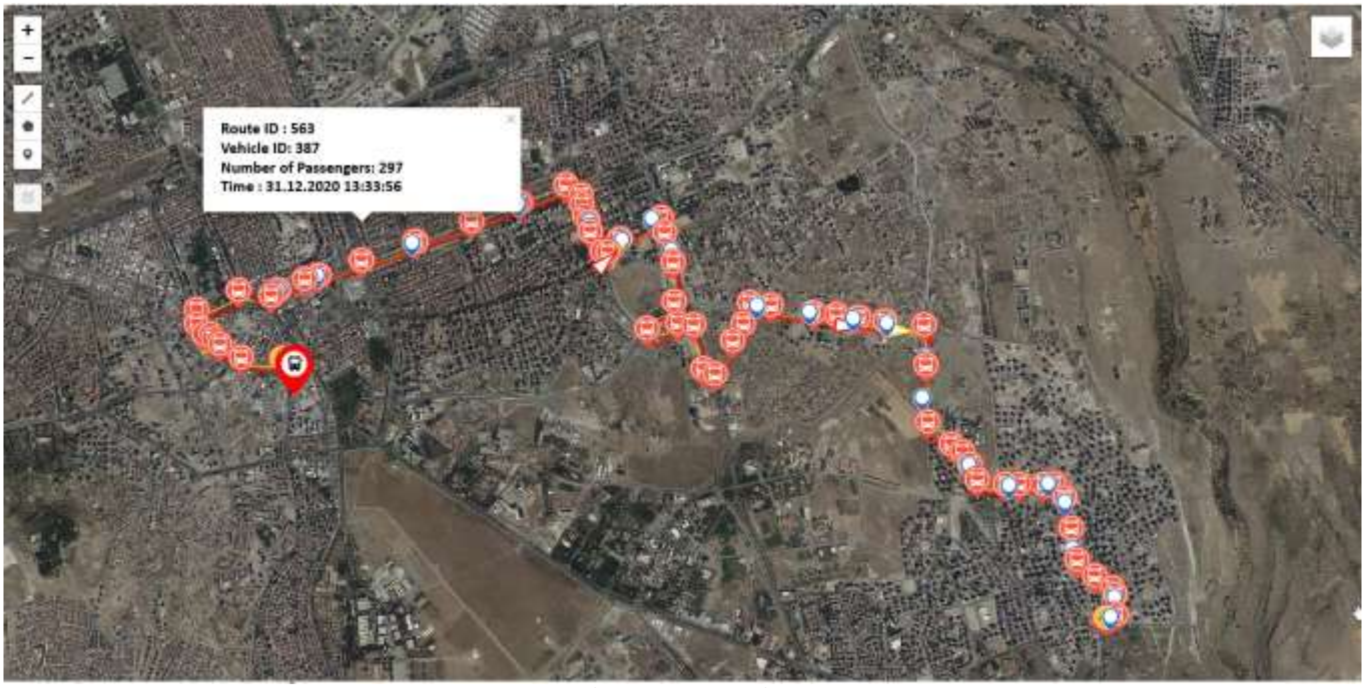


Fig.1. Display of vehicle location information on the map

- 🚌 : The map equivalent of the GPS position sent by the transportation vehicle
- 📍: Bus station
- —: Line route
- △ : Direction of line route

When the map layer and icons presented in Fig. 1 are examined in detail, it is seen that the bus with side number 387 is on the route of line 563. However, when it is desired to explore the accuracy and details of the voyage start and end times of the relevant vehicle, it is necessary to examine the locations it has sent before the route starts, instead of only the locations it sends along the route. In this case, the locations sent by the vehicle according to the vehicle side number and time information can be queried from the database system, and information about the vehicle can be obtained as a result of examining the relevant records on the map.

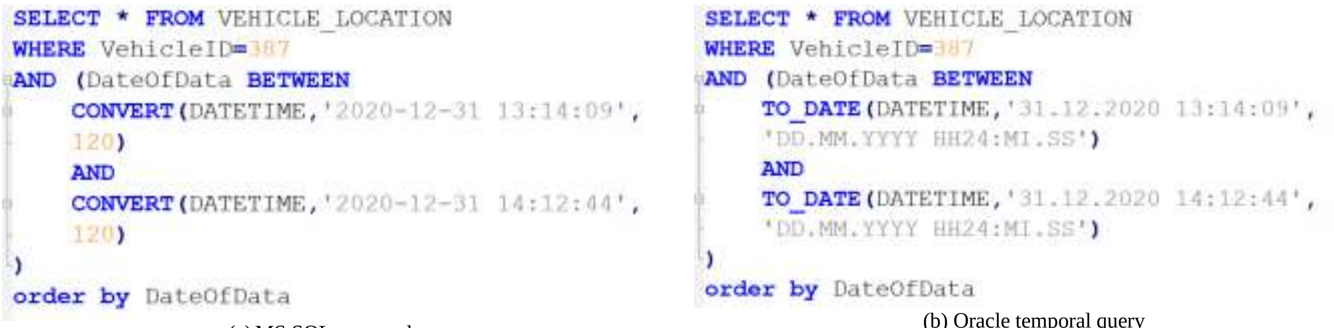
B. *Normalization Based Approach*

In database systems, the normalization process is defined as parsing a table with too many rows and columns into subsets of fewer rows and columns to eliminate repetitions [33, 34]. The structure of the database table called VEHICLE_LOCATION, which is created as a result of the Normalization process for the bus side number, line number, time information up to the second detail, point location information consisting of latitude and longitude values, the number of passengers in the vehicle and the speed information of the vehicle sent by the vehicles belonging to the KBB public transportation system. It is presented in Table I.

TABLE I
Information of VEHICLE_LOCATION Table

Column name	Column description	Data type
Vehicle ID	Side number to distinguish vehicles from each other	Integer
Route ID	Number of the route where the vehicle is running	Integer
Date of data	Sending time of GPS location information	Date time
Location	Point position of the vehicle (latitude and longitude)	Geometry (point)
Number of Passengers	Total number of passengers in the vehicle	Integer

VEHICLE_LOCATION table presented in Table I is a database table where the information sent instantly by all vehicles working in the public transportation system is stored. Queries run for analysis operations work on this table. For the voyage information presented in Fig. 1, the query sentences for MS SQL and Oracle database systems, which show all the information sent by the vehicle 10 minutes before and 10 minutes after the voyage, in chronological order and run in the table created with the normalization process, are shown in Fig. 2 (a) and 2 (b), respectively. In the query shown in Fig. 2 (a) and 2 (b), since the text expression to date field conversion functions of MS SQL and Oracle database systems and the parameters taken by the functions are different, the results of the queries are the same, but the query sentences differ from each other. In the MS SQL database system, the convert function is used to convert the text expression to date format, and the to date function is used in the Oracle database system [35]. In the MS SQL database system, the expression 120 in the convert function means “yyyy-MM-dd hh:mm:ss”.



```
SELECT * FROM VEHICLE_LOCATION
WHERE VehicleID=387
AND (DateOfData BETWEEN
    CONVERT(DATETIME,'2020-12-31 13:14:09',
    120)
    AND
    CONVERT(DATETIME,'2020-12-31 14:12:44',
    120)
)
order by DateOfData
```

(a) MS SQL temporal query

```
SELECT * FROM VEHICLE_LOCATION
WHERE VehicleID=387
AND (DateOfData BETWEEN
    TO_DATE(DATETIME,'31.12.2020 13:14:09',
    'DD.MM.YYYY HH24:MI.SS')
    AND
    TO_DATE(DATETIME,'31.12.2020 14:12:44',
    'DD.MM.YYYY HH24:MI.SS')
)
order by DateOfData
```

(b) Oracle temporal query

Fig.2. Temporal query

In the queries shown in Fig. 2 (a) and 2 (b), "yyyy" includes year information, "MM" shows month information, "dd" shows day information, "hh" - "HH24" show hour information, "mm" - "MI" show minute information, and "ss" shows seconds.

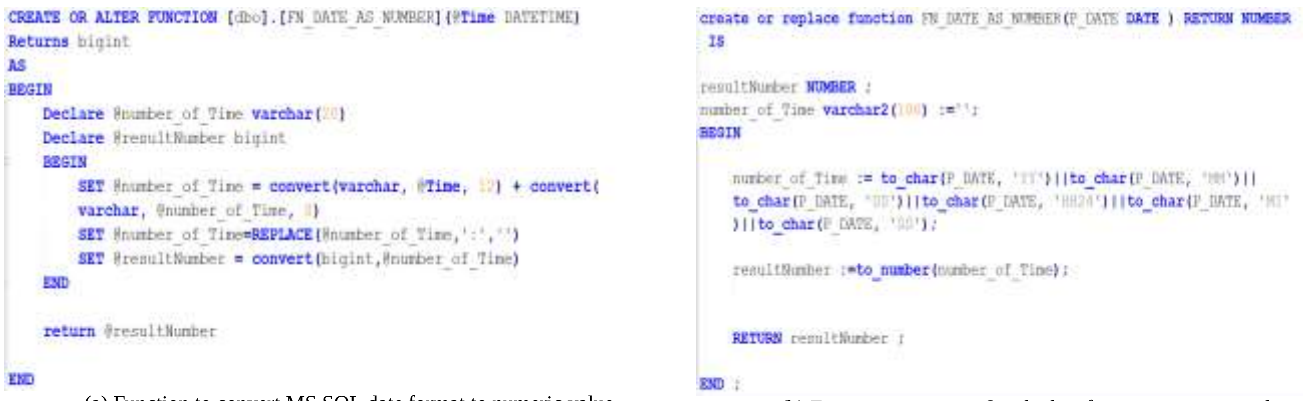
Although an index was created for the Vehicle ID and date fields of the VEHICLE_LOCATION table, which was created as a result of the normalization process in MS SQL and Oracle database systems, the running time of the queries presented in Fig. 2 (a) and 2 (b), which were run for the analysis of the public transport system, did not decrease at the desired level. Therefore, a novel approach has been proposed to speed up the running time of the query run with the normalization method and, therefore, to reduce the response time of the queries run for public transport system analysis operations.

C. Proposed Novel Approach

Accelerating the response time of the database by adding rows or columns that are not needed to a database system that has been normalized is defined as denormalization [22, 23]. As a result of the normalization processes, the database table described in detail in Table I was created, and the queries presented in Fig. 2 (a) and 2 (b) were run on this table. However, since the response time of the queries is very long, the analysis process takes longer. To overcome this problem, a new column of digit data type named Date_Number has been added to the database table presented in Table I as part of the denormalization process. In the new column called Date_Number, it is suggested to keep the numerical equivalent of the date-time data format in the Date field. To convert date fields to numeric values, the functions are shown in Fig. 3 (a) and 3 (b) created for MS SQL and Oracle databases systems, respectively.

When only a new record is added to VEHICLE_LOCATION table, the date field is converted to numeric values using these functions and recorded in the Date_Number field. The numeric value of the date field in the format YYMMDDHH24MISS was created by only two digits of the year information YY, the month information MM, the day information DD, the hour information HH24, the minute information MI and the second information SS. The query of the denormalization-based method is presented in Fig. 4.

The query presented in Fig. 4 shows the query format shown in Fig. 2 (a) and 2 (b) converted to MS SQL and Oracle databases after denormalization. In the query shown in Fig. 4, the Date_Number field created after the denormalization process is used instead of the date field. Filtering and sorting are done according to the Date_Number field, a number field, instead of a date field. As in the normalization method, an index is created for the Date_Number field in the denormalization-based method.



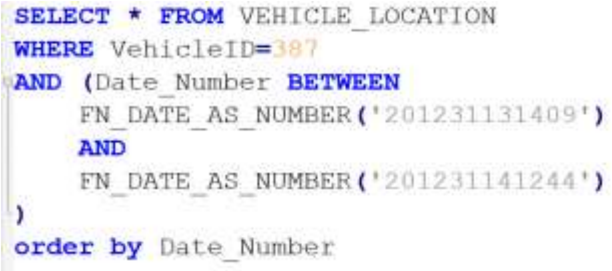
```
CREATE OR ALTER FUNCTION [dbo].[FN_DATE_AS_NUMBER](@Time DATETIME)
Returns bigint
AS
BEGIN
    Declare @number_of_Time varchar(20)
    Declare @resultNumber bigint
    BEGIN
        SET @number_of_Time = convert(varchar, @Time, 12) + convert(
        varchar, @number_of_Time, 8)
        SET @number_of_Time=REPLACE(@number_of_Time,':','')
        SET @resultNumber = convert(bigint,@number_of_Time)
    END

    return @resultNumber

END
```

(a) Function to convert MS SQL date format to numeric value

```
create or replace function FN_DATE_AS_NUMBER(P_DATE DATE ) RETURN NUMBER
 IS

resultNumber NUMBER ;
number_of_Time varchar2(100) :='';
BEGIN

    number_of_Time := to_char(P_DATE, 'YY')||to_char(P_DATE, 'MM')||
    to_char(P_DATE, 'DD')||to_char(P_DATE, 'HH24')||to_char(P_DATE, 'MI'
    )||to_char(P_DATE, 'SS');

    resultNumber :=to_number(number_of_Time);


    RETURN resultNumber ;

END ;
```

(b) Function to convert Oracle date format to numeric value

Fig.3. Function to convert date format to numeric value



```
SELECT * FROM VEHICLE_LOCATION
WHERE VehicleID=387
AND (Date_Number BETWEEN
     FN_DATE_AS_NUMBER('201231131409')
     AND
     FN_DATE_AS_NUMBER('201231141244')
)
order by Date_Number
```

Fig. 4. Query of the proposed method

IV. Experimental Evaluations

This study was carried out using the actual data of the vehicles working in the KBB public transportation system for December 2020. In Table II, the number of records of the entire data sets obtained from the KBB and the storage areas occupied by this data. In the KBB public transportation system, the communes sent by the vehicles for a month consist of 96,000,000 data sets. These data, which are sent instantly by hundreds of vehicles on a per-second basis, show big data characteristics because they are sent continuously and take up a lot of memory. Furthermore, since there is latitude and longitude information in the incoming data and it is known at which point the vehicle is on the geography, this data is also spatial data. Also, the data set is temporal since a vehicle sends data continuously at certain time intervals. Thus, the data set stores the historical data of the vehicles together with the time information.

TABLE II
DATA SETS

Sequence ID	Record Count	Size (megabyte)
1	100,000	551
2	250,000	1,380
3	500,000	2,765
4	1,000,000	5,546
5	2,000,000	11,136
6	3,000,000	16,815
7	4,000,000	22,622
8	5,000,000	28,595
9	10,000,000	57,647
10	15,000,000	87,508
11	20,000,000	118,428
12	25,000,000	150,640
13	30,000,000	184,383
14	40,000,000	250,638
15	50,000,000	320,316
16	60,000,000	393,989
17	70,000,000	472,260
18	80,000,000	555,784
19	90,000,000	645,265
20	96,000,000	712,803

Experimental evaluations sought answers to the following questions:

- What are the runtimes of normalization based and recommended approaches on incremental data set in MS SQL database?
- What are the runtimes of normalization based and recommended approaches on incremental data set in Oracle database?

To evaluate the performance of the methods in the experimental evaluations, the queries detailed in Fig. 2 (a), 2 (b) and 4 were run for 20 different vehicles operating in the KBB public transport system, and the average times were examined as the working time the methods. The experiments in this study were carried out on a desktop computer with Intel i7 11700 2.5GHz (8 Core), 32 GB Ram, 4 GB QUADRO graphics card, 2 TB SATA DISK, and Windows 10 Pro operating system installed.

A. Running Times of Methods in MS SQL Database System

In this experiment, the normalization and the result of the proposed method were examined in the MS SQL database system using the data sets detailed in Table II. In Table III, the working times of the methods are presented in seconds, and in Fig. 5, the operational times of the methods are examined in minutes. As the number of records in the data set increases, the working time of the methods increases.

TABLE III
RUN TIMES OF METHODS IN MS SQL DATABASE SYSTEM

Sequence ID	Normalization Based Method (millisecond)	Proposed Method (millisecond)
1	3	3
2	5	4
3	12	10
4	28	15
5	64	26
6	90	32
7	161	64
8	254	118
9	584	240
10	874	270
11	1,175	310
12	1,574	340
13	1,807	390
14	2,400	430
15	3,101	466
16	3,751	524
17	4,400	572
18	5,027	604
19	5,913	722
20	6,215	809

When Table III and Fig. 5 are examined in detail, there is almost no difference between normalization and denormalization methods up to 1,000,000 data sets. However, in cases where the data set has more than 10,000,000 rows of records, the response time of the query run with the normalization method exceeds a second. If the size of the data set is 96,000,000, the response time of the normalization-based method is over 6 seconds, which is a very long time for a system that works in real-time and is analyzed instantly. On the other hand, as the size of the data set increases, the working time of the normalization-based method increases much more than the proposed method. Although the proposed method's performance increases as the data set's size increases, a linear increase cannot be observed. The proposed method can run analysis queries in a very short time frame, such as 0.8 seconds, even in the worst case where the data set has

96,000,000 records. When the queries shown in Fig. 2 (a) and Fig. 4 are examined in detail, not only filtering but also sorting is performed. In the denormalization-based method, although it is not needed in the database design, the Date_Number column is added. The data in the Date column is converted to number format with the function presented in Fig. 3 (a). This column, which was added as part of the denormalization method, accelerated the performance of the database by approximately eight times for 96,000,000 data sets.

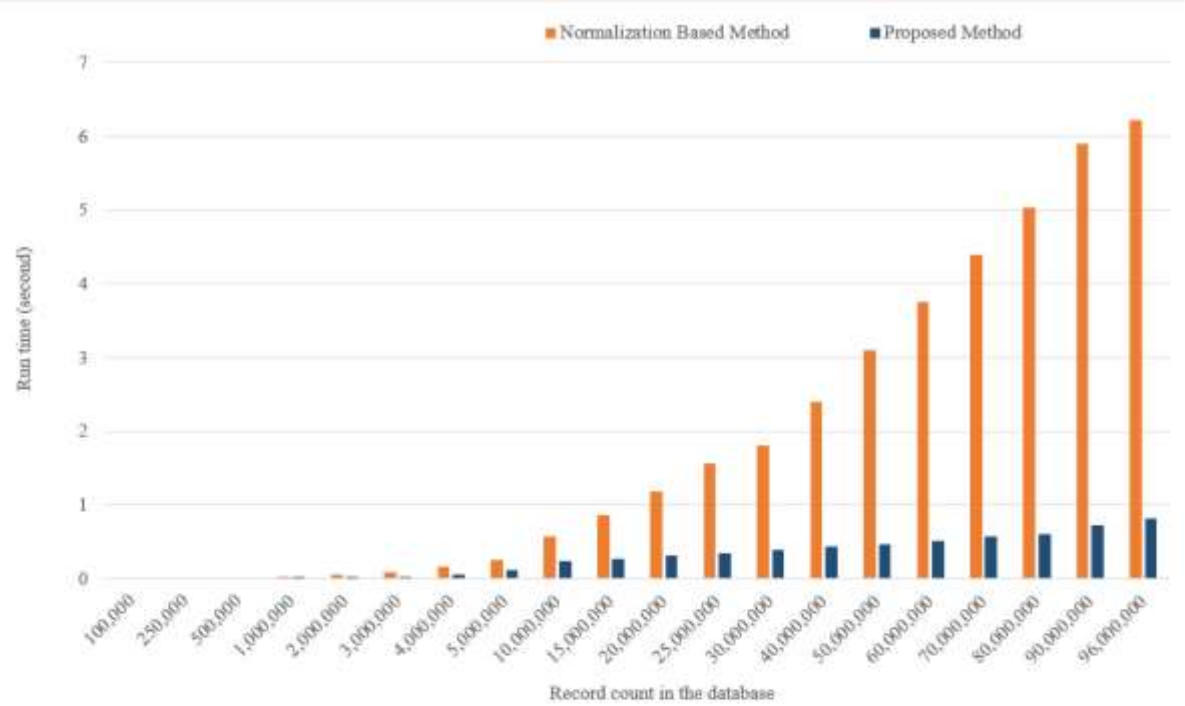


Fig. 5. Run times of methods in MS SQL database system

B. *Running Times of Methods in Oracle Database System*

In this experiment, the normalization and the result of the proposed denormalization-based method were examined in the Oracle database system using the data sets detailed in Table II. In Table IV, the working times of the methods are presented in seconds, and in Fig. 6, the working times of the methods are examined in minutes. As in the MS SQL database system, the operating times of the methods increase as the number of records in the data set increases in the Oracle database system. Similar to the MS SQL database system, when Table IV and Fig. 6 are examined in detail, there is almost no difference between normalization and denormalization methods up to 1,000,000 data sets. However, unlike the MS SQL database system, the response time of the query executed with the normalization method exceeds a second when the data set has over 20,000,000 rows of records. If the data set size is 96,000,000, the response time of the normalization-based method is over 5 seconds, which is a very long time for a system that works in real-time and is analyzed instantly. Similar to the experiments performed in the MS SQL database system, as the data set's size increases, the normalization-based method's running time increases much more than the denormalization-based method. In the Oracle database system, although the performance of the proposed denormalization-based method increases as the size of the data set increases, a linear increase cannot be observed. The proposed method can run analysis queries in a very short time frame, such as 0.7 seconds, even in the worst case where the data set has 96,000,000 records. In the Oracle database system, the Date_Number column added within the scope of the denormalization method accelerated the performance of the database to 96,000,000 data sets, approximately eight times as in the MS SQL database system.

TABLE IV
RUN TIMES OF METHODS IN ORACLE DATABASE SYSTEM

Sequence ID	Normalization Based Method (millisecond)	Proposed Method (millisecond)
1	2	2
2	4	3
3	9	8
4	24	14
5	55	21
6	80	31
7	150	39
8	270	52
9	500	60
10	810	95
11	1,030	120
12	1,300	150
13	1,600	190
14	2,150	250
15	2,700	310
16	3,300	380
17	3,800	430
18	4,300	490
19	4,980	588
20	5,311	658

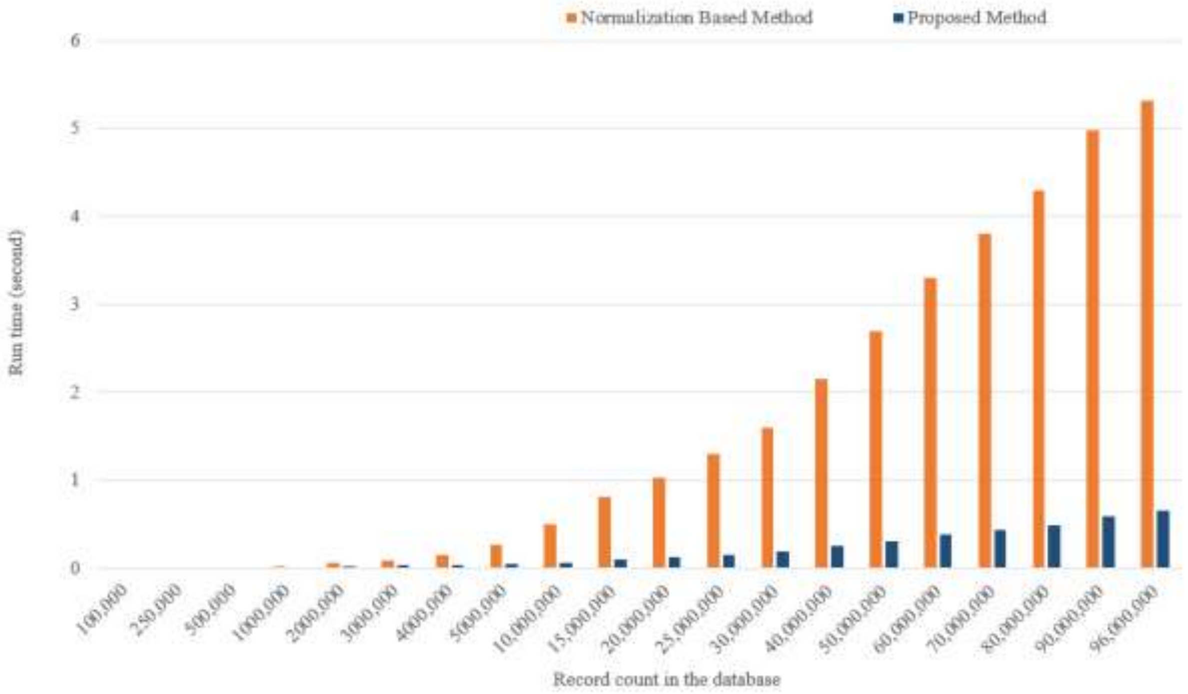


Fig. 6. Run times of methods in Oracle database system

C. Application

The application developed within the scope of this study is actively used by the KBB. It is used for route violation, stop violation, and control of the journey time of the lines operating in the public transportation system. Thanks to the proposed application, the list of lines that were violated is automatically presented, and by clicking the detail button, how the violation was made is displayed on the map thanks to the proposed application. Before the proposed practice was used, the detection of violations was only detected upon complaint. Thanks to the application, the detection of violations can be viewed instantly. The screenshot of the application used is presented in Fig. 7 and Fig. 8 in the Appendix part. Fig. 7 shows the list of voyages operating according to important criteria such as line name, number, start time, and end time. Fig. 8 presents the list of violating lines. In Fig. 8, the word GI indicates route violation, the word DI indicates to stop the violation, and the word HD indicates the line number changes while the voyage continues.

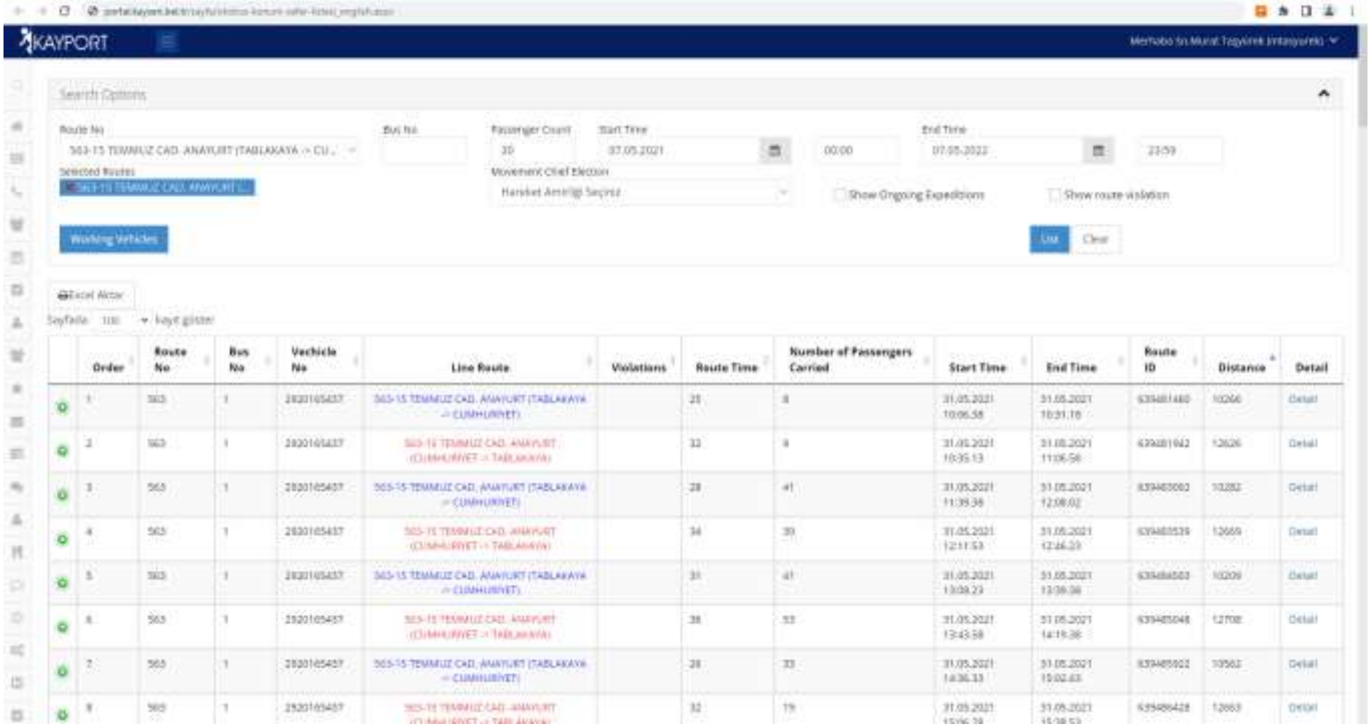
V. Conclusion

Big data systems can be used in transportation, banking, communication, media, entertainment, healthcare, education, manufacturing, etc. It is widely used in many different fields. Due to big data as the oil, the low latency of NoSQL database systems and the additional cost of data stored in relational database systems, and showing big data characteristics to OLAP systems are big problems. This study proposes a new approach to improve date-based filtering and sorting performance in relational database systems where large volumes of spatial data containing time information are stored. The performance of the proposed method is examined using actual data in MS SQL and Oracle database systems. Experimental evaluations Since normalization and denormalization-based methods work very closely in data sets with less than 1,000,000 record numbers. There is no need for a denormalization-based approach in such data sets. However, as the amount of data in the data set increases, the working time of the normalization-based method increases more than the denormalization-based method. When the data set with 96,000,000 records is examined, the denormalization-based method responded eight times faster than the normalization-based method, since it performs both filtering and sorting operations numerically. Experimental evaluations have shown that as the number of records in the data set increases, the denormalization-based method works much faster than the normalization-based method.

Acknowledgment

We would like to thank the KBB for sharing large volumes of spatial data containing real-time information of the vehicles operating in the public transportation system used within the scope of this study. In addition, this study was supported by the Scientific Research Projects Coordination Unit of Kayseri University within the scope of project #FHD-2021-1043.

APPENDIX

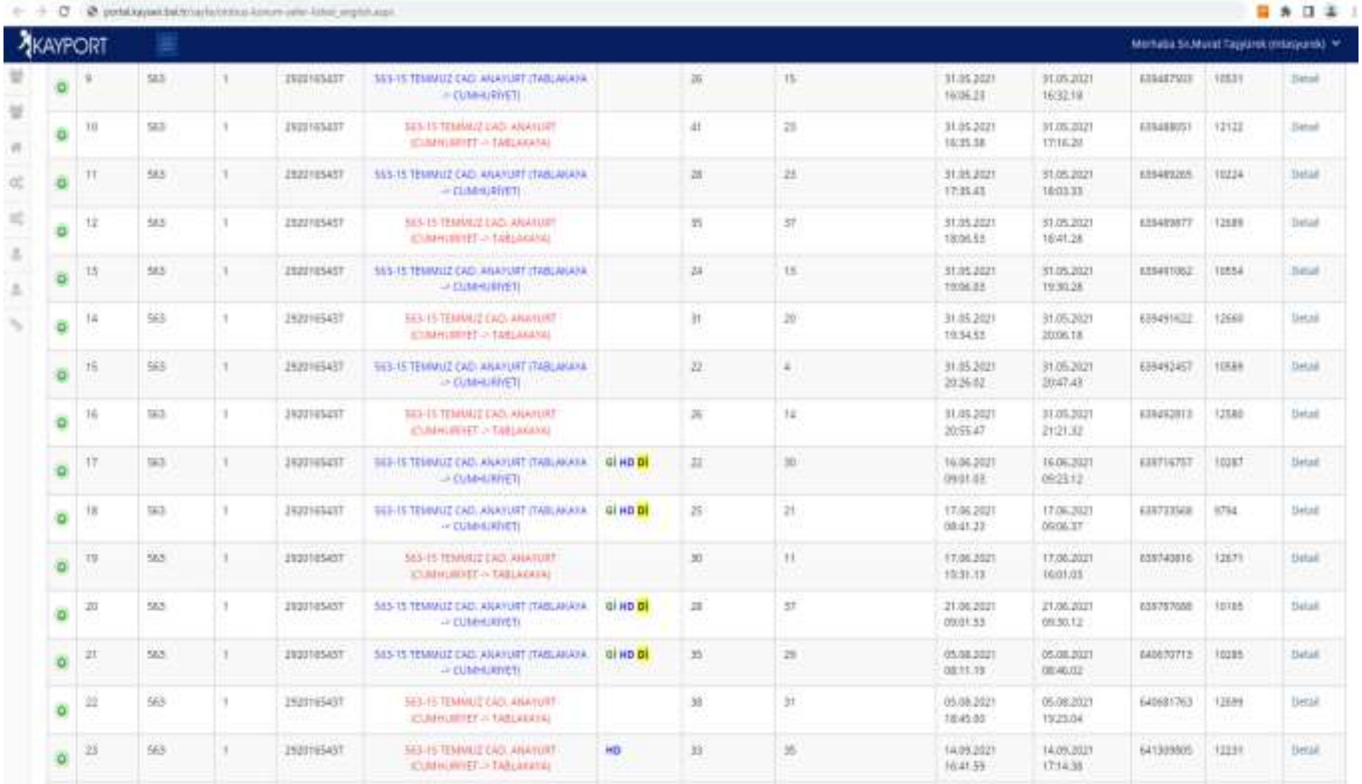


Route No: 563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CU...
Bus No:
Passenger Count: 30
Start Time: 07.05.2021 00:00
End Time: 07.05.2022 23:59
Selected Routes: 563-15 TEMMUZ CAD. ANAYURT (...
Movement Chief Election: Hareket Amirliği Seçiniz
□ Show Ongoing Expeditions
□ Show route violation

Sayfada 100 kayıt göster

	Order	Route No	Bus No	Vechicle No	Line Route	Violations	Route Time	Number of Passengers Carried	Start Time	End Time	Route ID	Distance	Detail
●	1	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)		25	8	31.05.2021 10:06:58	31.05.2021 10:31:18	639481480	10266	Detail
●	2	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)		32	9	31.05.2021 10:35:13	31.05.2021 11:06:58	639481942	12626	Detail
●	3	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)		28	41	31.05.2021 11:39:38	31.05.2021 12:08:02	639483082	10282	Detail
●	4	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)		34	39	31.05.2021 12:11:53	31.05.2021 12:46:23	639483539	12669	Detail
●	5	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)		31	41	31.05.2021 13:08:23	31.05.2021 13:38:38	639484583	10209	Detail
●	6	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)		38	53	31.05.2021 13:43:58	31.05.2021 14:19:38	639485048	12708	Detail
●	7	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)		28	33	31.05.2021 14:36:33	31.05.2021 15:02:43	639485922	10562	Detail
●	8	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)		32	19	31.05.2021 15:06:28	31.05.2021 15:38:52	639486428	12663	Detail

Fig. 7. Search Screen Used by KBB



	Order	Route No	Bus No	Vechicle No	Line Route	Violations	Route Time	Number of Passengers Carried	Start Time	End Time	Route ID	Distance	Detail
●	9	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)		26	15	31.05.2021 16:06:23	31.05.2021 16:32:18	639487503	10531	Detail
●	10	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)		41	25	31.05.2021 16:35:38	31.05.2021 17:16:20	639488051	12122	Detail
●	11	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)		28	25	31.05.2021 17:35:43	31.05.2021 18:03:33	639489265	10224	Detail
●	12	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)		35	57	31.05.2021 18:06:53	31.05.2021 18:41:28	639489877	12589	Detail
●	13	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)		24	15	31.05.2021 19:06:33	31.05.2021 19:30:28	639491062	10554	Detail
●	14	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)		31	20	31.05.2021 19:34:53	31.05.2021 20:06:18	639491622	12560	Detail
●	15	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)		22	4	31.05.2021 20:26:02	31.05.2021 20:47:43	639492457	10589	Detail
●	16	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)		26	14	31.05.2021 20:55:47	31.05.2021 21:21:32	639492813	12580	Detail
●	17	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)	Gİ HD Dİ	22	30	16.06.2021 09:01:03	16.06.2021 09:23:12	639716757	10287	Detail
●	18	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)	Gİ HD Dİ	25	21	17.06.2021 08:41:23	17.06.2021 09:06:37	639733568	9794	Detail
●	19	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)		30	11	17.06.2021 15:31:13	17.06.2021 16:01:03	639743816	12671	Detail
●	20	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)	Gİ HD Dİ	28	57	21.06.2021 09:01:53	21.06.2021 09:30:12	639757088	10185	Detail
●	21	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (TABLAKAYA -> CUMHURIYET)	Gİ HD Dİ	35	29	05.08.2021 08:11:19	05.08.2021 08:46:02	640670713	10285	Detail
●	22	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)		38	31	05.08.2021 18:45:00	05.08.2021 19:23:04	640681763	12689	Detail
●	23	563	1	2920165437	563-15 TEMMUZ CAD. ANAYURT (CUMHURIYET -> TABLAKAYA)	HD	33	35	14.09.2021 16:41:59	14.09.2021 17:14:38	641309905	12231	Detail

Fig. 8. List of Violated Routes Used by KBB

REFERENCES

[1] P. K. Malik, R. Sharma, R. Singh, A. Gehlot, S. C. Satapathy, W. S. Alnumay, D. Pelusi, U. Ghosh, and J. Nayak, "Industrial internet of things and its applications in industry 4.0: State of the art," Computer Communications, vol. 166, pp. 125–139, 2021.

[2] V. Suma et al., "Internet-of-things (iot) based smart agriculture in indiaan overview," Journal of ISMAC, vol. 3, no. 01, pp. 1–15, 2021.

[3] M. Ghasemaghaei, "Understanding the impact of big data on firm performance: The necessity of conceptually differentiating among big data characteristics," International Journal of Information Management, vol. 57, p. 102055, 2021.

[4] C. Fan, D. Yan, F. Xiao, A. Li, J. An, and X. Kang, "Advanced data analytics for enhancing building performances: From data-driven to big

data-driven approaches," in Building Simulation, vol. 14, no. 1. Springer, 2021, pp. 3–24.

[5] M. Naeem, T. Jamal, J. Diaz-Martinez, S. A. Butt, N. Montesano, M. I. Tariq, E. De-la Hoz-Franco, and E. De-La-Hoz-Valdiris, "Trends and future perspective challenges in big data," in Advances in Intelligent Data Analysis and Applications. Springer, 2022, pp. 309–325.

[6] J. Ranjan and C. Foropon, "Big data analytics in building the competitive intelligence of organizations," International Journal of Information Management, vol. 56, p. 102231, 2021.

[7] M. L. Larrea and D. K. Urribarri, "Visualization technique for comparison of time-based large data sets," in Conference on Cloud Computing, Big Data & Emerging Topics. Springer, 2021, pp. 179–187.

[8] J. D. Dinneen and C. Brauner, "Information-not-thing: further problems with and alternatives to the belief that information is physical," 2017.

[9] M. Vaitis, H. Feidas, P. Symeonidis, V. Kopsachilis, D. Dalaperas, N. Koukourouvli, D. Simos, and S. Taskaris, "Development of a spatial database and web-gis for the climate of greece," Earth Science Informatics, vol. 12, no. 1, pp. 97–115, 2019.

[10] M. Amin, G. W. Romney, P. Dey, and B. Sinha, "Teaching relational database normalization in an innovative way," Journal of Computing Sciences in Colleges, vol. 35, no. 2, pp. 48–56, 2019.

[11] S. Alqithami, "A serious-gamification blueprint towards a normalized attention," Brain Informatics, vol. 8, no. 1, pp. 1–13, 2021.

[12] I. Oditis, Z. Bicevska, J. Bicevskis, and G. Karnitis, "Implementation of nosql-based data wareh," Baltic Journal of Modern Computing, vol. 6, no. 1, pp. 45–55, 2018.

[13] I. Hrubaru, G. Talab˘a, and M. Fotache, "A basic testbed for json data processing in sql data servers," in Proceedings of the 20th International Conference on Computer Systems and Technologies, 2019, pp. 278–283.

[14] Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp et al., "Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019," Journal of Chemical & Engineering Data, vol. 64, no. 12, pp. 5985–5998, 2019.

[15] P. Bouros and N. Mamoulis, "Spatial joins: what's next?" SIGSPATIAL Special, vol. 11, no. 1, pp. 13–21, 2019.

[16] V. K. Myalapalli, T. P. Totakura, and S. Geloth, "Augmenting database performance via sql tuning," in 2015 International Conference on Energy Systems and Applications. IEEE, 2015, pp. 13–18.

[17] W. G. Pedrozo and M. S. M. G. Vaz, "A tool for automatic index selection in database management systems," in 2014 International Symposium on Computer, Consumer and Control. IEEE, 2014, pp.1061–1064.

[18] J. Correia, M. Y. Santos, C. Costa, and C. Andrade, "Fast online analytical processing for big data warehousing," in 2018 International Conference on Intelligent Systems (IS). IEEE, 2018, pp. 435–442.

[19] H. Sulistiani, S. Setiawansyah, and D. Darwis, "Penerapan metode agile untuk pengembangan online analytical processing (olap) pada data penjualan (studi kasus: Cv adilia lestari)," Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi, vol. 6, no. 1, pp. 50–56, 2020.

[20] U. Erdinc, H. N. Bulus, and C. Erdoğan, "Veritabanı tasarımının yazılım performansına etkisi: Normalizasyona karşı denormalizasyon," Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, vol. 22, no. 2, pp. 887–895, 2018.

[21] B. Alshemaimri, R. Elmasri, T. Alsahfi, and M. Almotairi, "A survey of problematic database code fragments in software systems," Engineering Reports, vol. 3, no. 10, p. e12441, 2021.

[22] D. Milicev, "Hyper-relations: A model for denormalization of transactional relational databases," IEEE Transactions on Knowledge and Data Engineering, 2021.

[23] I. N. Chaparro-Cruz and J. A. Montoya-Zegarra, "Borde: Boundary and sub-region denormalization for semantic brain image synthesis," in 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, 2021, pp. 81–88.

[24] R. L. d. C. Costa, J. Moreira, P. Pintor, V. dos Santos, and S. Lifschitz, "A survey on data-driven performance tuning for big data analytics platforms," Big Data Research, vol. 25, p. 100206, 2021.

[25] A. H. Chill´on, D. S. Ruiz, and J. G. Molina, "Towards a taxonomy of schema changes for nosql databases: the orion language," in International Conference on Conceptual Modeling. Springer, 2021, pp. 176–185.

[26] E. Gupta, S. Sural, J. Vaidya, and V. Atluri, "Attribute-based access control for nosql databases," in Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy, 2021, pp. 317–319.

[27] J. Yang, Y. Yue, and K. Rashmi, "A large-scale analysis of hundreds of in-memory key-value cache clusters at twitter," ACM Transactions on Storage (TOS), vol. 17, no. 3, pp. 1–35, 2021.

[28] A. Hillenbrand, U. St¨orl, S. Nabiyev, and M. Klettke, "Self-adapting data migration in the context of schema evolution in nosql databases," Distributed and Parallel Databases, pp. 1–21, 2021.

[29] A. Rafique, D. Van Landuyt, E. H. Beni, B. Lagaisse, and W. Joosen, "Cryptdice: Distributed data protection system for secure cloud data storage and computation," Information Systems, vol. 96, p. 101671, 2021.

[30] B. Vela, J. M. Cavero, P. C´aceres, A. Sierra, and C. E. Cuesta, "Defining a nosql document database of accessible transport routes," in 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, 2017, pp. 1125–1129.

[31] S. P. R. Asaithambi, R. Venkatraman, and S. Venkatraman, "Mobda: Microservice-oriented big data architecture for smart city transport systems," Big Data and Cognitive Computing, vol. 4, no. 3, p. 17, 2020.

[32] M. T. Gonz´alez-Aparicio, M. Younas, J. Tuya, and R. Casado, "Testing of transactional services in nosql key-value databases," Future Generation Computer Systems, vol. 80, pp. 384–399, 2018.

[33] J. S. Fong, Information Systems Reengineering, Integration and Normalization: Heterogeneous Database Connectivity. Springer Nature, 2021.

[34] M. Taşyürek, "Mekansal verilerin sıklıkla güncellendiği coğrafi bilgi sistemleri arama işleminde denormalizasyon yöntemi," Avrupa Bilim ve Teknoloji Dergisi, no. 24, pp. 18–23, 2021.

[35] J. Rand and A. Miranskyy, "On automatic parsing of log records," in 2021 IEEE/ACM 43rd International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER). IEEE, 2021, pp. 41–45.

BIOGRAPHIES

MURAT TASYUREK was born in Turkey in 1986. He received the B.S. degree, the M.Sc. degree, and the Ph.D. degree in computer engineering from Erciyes University in 2009, 2011 and 2020, respectively. He worked as a software developer computer engineer for the Kayseri metropolitan municipality between 2010 and 2021. He has been working as an assistant prof in the computer engineering department of Kayseri University since 2021. His research interests include image registration, data mining, GIS (Geographic Information Systems) analysis, big data, deep learning, and optimization algorithms.

Actionable Data Visualization for Air Quality Data in the Istanbul Location

Damla Mengüş and Bihter Daş

***Abstract*— Air pollution is increasing day by day due to the increasing population, urbanization, and industrial development. In our country, the amounts of pollutants in the air are recorded every day at different points. These recorded data continue to be collected in an increasing amount day by day. Information overload, which renders the data meaningless, complicates the interpretation of these data. One of the ways to solve this problem is to visualize curves and trends in measured pollution concentrations over time. In this study, using the data provided by the continuous monitoring center of the Turkey Ministry of Environment, Urbanization and Climate Change, visualization of different pollutants in the air was provided. Scatter plots, line scatter plots, and bar plots were used as data visualization. Data visualization makes it easy for non-experts to estimate air quality information from the concentration profiles displayed.**

Index Terms— **Data visualization, air pollution, air quality assessment, visual analytics**

I. INTRODUCTION

AIR POLLUTION is defined as the presence of foreign substances above normal, which should not be present in the air, which adversely affect human health and environmental balance. There are many factors to air pollution, especially the increasing population, urbanization, and industrial developments [1-3]. Carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2), ozone (O3), particulate matter (PM), and NOx, which is a combination of nitric oxide and nitrogen dioxide, are the leading gases that cause air pollution[4-6]. Particulate matter PM10 and PM2.5 are one of the most important pollutants affecting human health. Since it is very small in size, it passes through our respiratory system very easily and penetrates our lungs. May cause cancer if inhaled for a long time. As other pollutants can easily penetrate the lungs, all of them are very dangerous for human health [7-9]. Evaluation of air quality is made according to the air quality index. This index is a measure used to express air quality. As the measured value gets larger, it is understood that the air starts to have negative effects on human health. Air quality index 0-50 range is good, 51-100 range is medium, 101-150 range is sensitive, 151-200 range is unhealthy, 201-300 range is bad, 301-500 range is dangerous. There is a high probability of experiencing health problems with a value of 151 and above, and it is not necessary to go to the open area during these times [10,11].

In this study, data visualization is carried out to observe the air quality by using the air pollutant data of the Basaksehir district of Istanbul city. There are continuous monitoring centers (SIMs) at 39 different points in Istanbul. It is very suitable for monitoring the air quality in different parts of the city and making inferences from the data. The differences in the districts of Istanbul, located in the middle or by the sea, provide better observation with visualization tools. Figure 1 shows the location of the Başakşehir district.

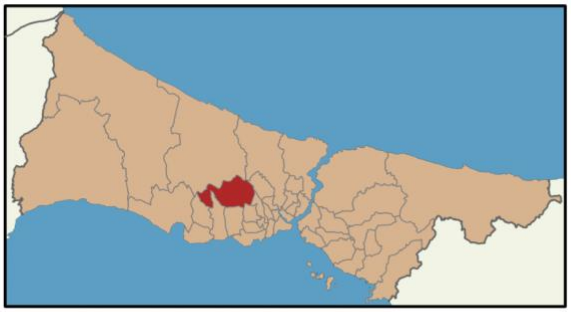


Fig.1. The location of the Basaksehir

A. Contributions of the paper

The main contributions of the study are listed below.

1) Air quality indices (AQI) and environmental data are combined with data visualization.
2) A practical experiment on which visualization tools would be appropriate for what type of data is provided.
3) Visualization methods help non-experts to interpret big data.

The rest of the article is organized as follows. The methods used are mentioned in Chapter 2. In Chapter 3, the application steps and the data used are mentioned. In Chapter 4, the results of the study are mentioned. In section 5, inferences are made.

Damla MENGÜŞ is with the Department of Computer Engineering, Technology Faculty, Marmara University, İstanbul TÜRKİYE e-mail: damla.mengus@gmail.com

Bihter DAŞ is with the department of Software Engineering, Technology Faculty, Firat University, Elazig TÜRKİYE e-mail: bihterdas@firat.edu.tr

Manuscript received September 27, 2022; accepted October 21, 2022.
DOI: 10.17694/bajece.1180676

B. *Related works*

There are many areas in the literature where data visualization is used. In most places with big data, visualization is used to understand the data[12,13]. Bachechi et al. used information visualization techniques to analyze urban traffic data and the effect of traffic emissions on urban air quality. Traffic data statistics for months or years provided a clear understanding of the similarities and differences between days[14]. Von Bromssen et al. studied the acidification of the river by using Swedish riverbank data during 1988–2017. They have tried to summarize complex information over nearly thirty years of data and have used data visualization while doing this [15]. Huang et al. Shenzhen, a mega-city in China, has made efforts to promote the transition to green transport by enforcing license plate restrictions. However, it is unclear whether the restrictions improve urban air quality. They have studied the effect of diesel vehicles on air pollution. Thanks to the data visualization tools used in the study, the result was easily achieved [16]. Pérez-Campuzano et al. took 30 years of data from 18 different US passenger airlines and made visualization on these data in their study. As a result of the study, it was revealed how much the passenger airline of the USA was affected during the Covid-19 period [17]. In the study of Prasad et al., existing data visualization methods have been enhanced by spectral modeling to overcome the problem of cluster bias on non-CS datasets, which efficiently recognizes the spectral features of non-CS datasets and cluster patterns[18]. New visualization techniques are used not only for environmental data, but also for other types of data such as satellite information, X-ray spectra processing or big data [19-22]. For this purpose, a software platform was developed with an integration in the form of two measurement stations and satellite information in a unified view to process publicly available data from various sources. In [23], authors superposed health risk information from 9 different Air Quality Indices (AQIs) on different kinds of graphs. They visualizated tha data, which is obtained from two monitoring stations located in regions, Ghent and Vielsalm in Belgian Environment Agency.

II. BACKGROUND

In this part of the study, general information about data visualization methods and incomplete data completion techniques is given. The process of making large and complex data meaningful and understandable with certain graphics is called data visualization. Three different types of charts were used in this study:

Point scatter plot: Point scatter plot is a data visualization method created using two different numerical data. It allows us to directly see the connection between these two numerical values. We can visualize not only two values, but many values at the same time through different colors or sizes. Thus, complex data becomes more understandable thanks to visualization.

Line scatter plot: Line scatter plot gives similar output when used in the same way as point scatter plot. However, the line scatter plot is easier to read than the point scatter plot in some cases. Breakpoints that are not clearly visible in the point scatter plot can be observed better with the line scatter plot. For this reason, data visualization with line scatter plot may be more advantageous depending on the usage area.

Bar plot: It provides the opportunity to visualize data in categories with bar plot. Data can be displayed in multiple groups in a single image. This provides an advantage for the bar plot.

One of the other problems when working with data is the problem of missing data. Accurate inference or visualization cannot be made due to missing data. In such cases, there are various algorithms that can be used to complete the missing data. Two different methods were used for this study. Based on the results, it was decided that the most appropriate method for this study was to complete the missing data with the mean technique.

Complete missing data with k-nn technique: K-nearest neighbor (KNN) is a kind of algorithm used for classification and regression in supervised learning. Training and testing are pretty much the same. It is not an ideal algorithm to complete missing values in large datasets.

Complete missing data with mean technique: In this method, the missing data part is filled by taking the average of the other data in the area where the missing data is located. Data range is very important for using this method. Using this method in data sets with very high data range increases the error rate.

III. METHODOLOGY AND IMPLEMENTATION

In this study, firstly annual data were collected and then missing data was completed with k-nn and average algorithms. Then, different visualization methods were applied to the data. Point scatter plot, line scatter plot and bar plot were used on the data, respectively. The mean technique was used as the missing data completion algorithm in the study because the data range is low, using mean in such data provides better performance. The flowchart showing the application steps is shown in Figure 2.

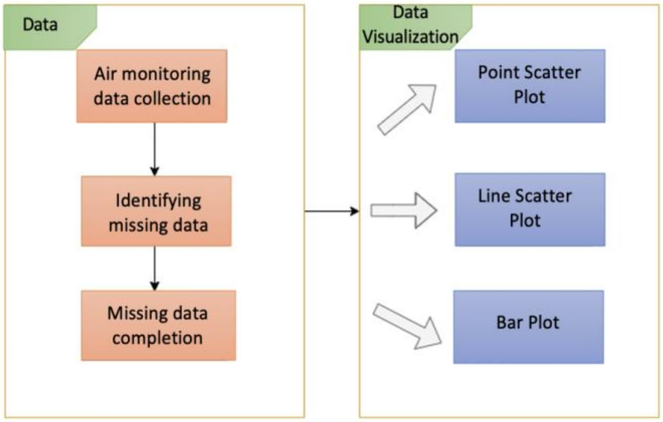
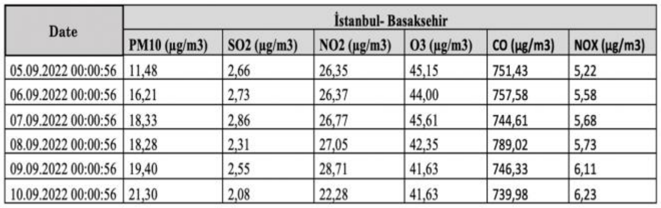


Fig. 2. The flowchart of the study

A. *Data description*

In this study, the data provided by the continuous monitoring center of the Turkey Ministry of Environment, Urbanization and Climate Change were used [24]. It consists of data on different pollutants recorded daily or hourly. In our study, visualization was made on the data of six pollutants for one year. These inhibitors are NOx, which is a combination of carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2), ozone (O3), particulate matter (PM), nitric oxide and nitrogen dioxide. Table 1 shows the first six rows of the data set.

Table 1. First six rows of data set



Date	İstanbul- Basaksehir					
	PM10 (µg/m3)	SO2 (µg/m3)	NO2 (µg/m3)	O3 (µg/m3)	CO (µg/m3)	NOX (µg/m3)
05.09.2022 00:00:56	11,48	2,66	26,35	45,15	751,43	5,22
06.09.2022 00:00:56	16,21	2,73	26,37	44,00	757,58	5,58
07.09.2022 00:00:56	18,33	2,86	26,77	45,61	744,61	5,68
08.09.2022 00:00:56	18,28	2,31	27,05	42,35	789,02	5,73
09.09.2022 00:00:56	19,40	2,55	28,71	41,63	746,33	6,11
10.09.2022 00:00:56	21,30	2,08	22,28	41,63	739,98	6,23

B. *Data preprocessing*

In this study, the average method was used while filling the missing values of the data set consisting of one-year data. The reason for choosing the incomplete data completion algorithm with the mean is that the data set is large and the data range is small.

C. *Actionable data visualization*

The most important reason for choosing point scatter, line scatter and bar plot as visualization methods in this study is that the data set is independent of each other. While simpler visualization tools are used in independent and unrelated data sets, different visualization methods are used in linked related data.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The visualizations that emerged as a result of the study are given below, respectively. In Figure 3, it is seen that the PM10 value is not very high throughout the year and is high on very rare dates. In particular, it is clearly seen that it reached the highest point in April 2022.

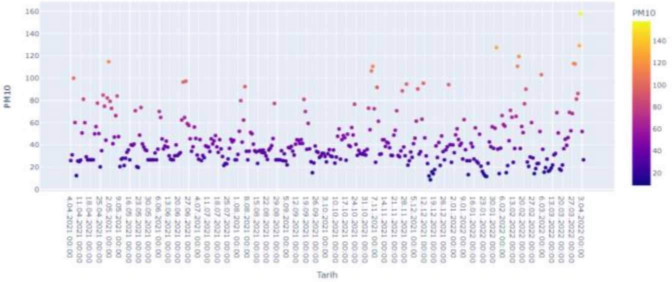


Fig. 3. Point Scatter for PM10

In Figure 4, it is seen that the SO2 value is much higher, especially in April and May, compared to the rest of the year. It can be said that the annual sulfur dioxide value is higher than normal.

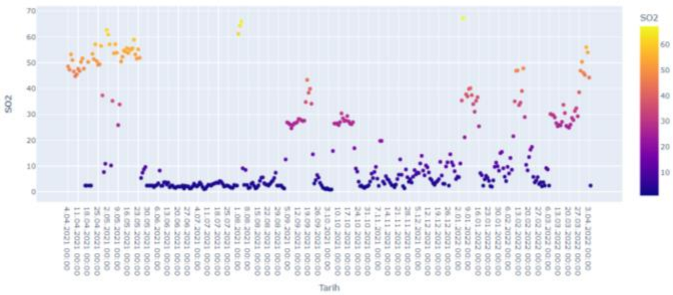


Fig. 4. Point Scatter for SO2

In Figure 5, annual NO2 values are more evenly distributed. High values can be easily seen.

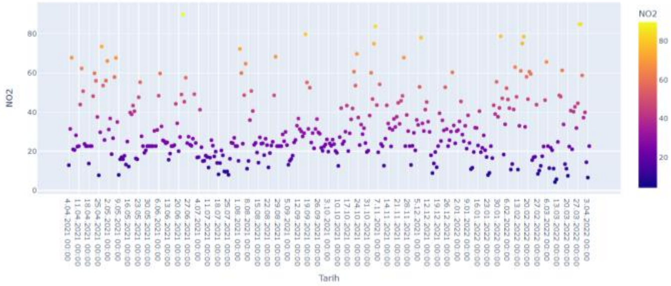


Fig. 5. Point Scatter for NO2

Figure 6 shows a more balanced distribution for the NOX value. It is observed that the annual values are normal and very slightly increase to high values.

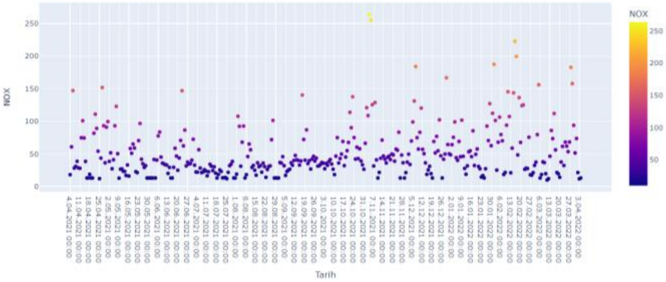


Fig. 6. Point Scatter for NOX

In Figure 7, very high O3 values were observed in May, July and August. Even looking at the annual value in general, it is mostly seen to be high.

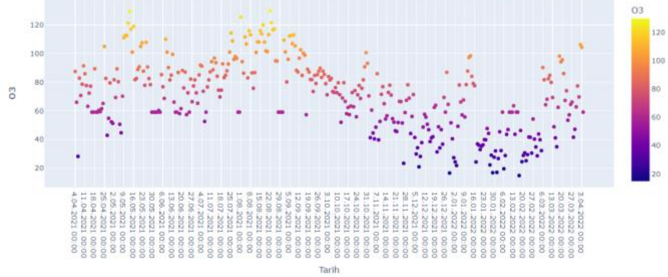


Fig. 7. Point Scatter for O3

The visualizations for the line scatter on the same data set are given below, respectively. In the Figure 8-12, the peaks are more prominent. The highest and lowest values can be easily observed in these figures. In Figure 8, unlike the point scatter, the highest PM10 value was shown in March. The upper values are more prominent in the image.

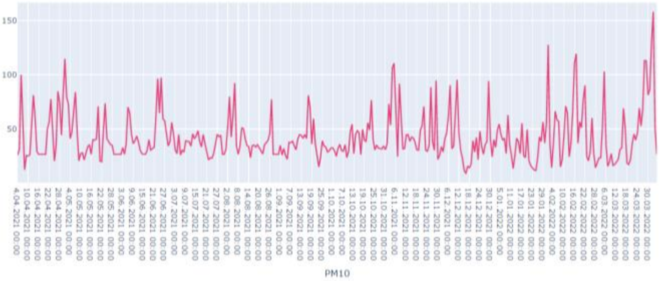


Fig. 8. Line Scatter for PM10

Figure 9 shows that the values are distributed unevenly. High and low values are easily visible.

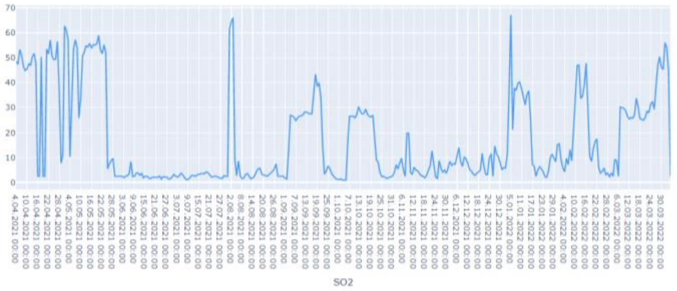


Fig. 9. Line Scatter for SO2

Figure 10 shows that the NO2 value is evenly distributed. High and low values are evident.

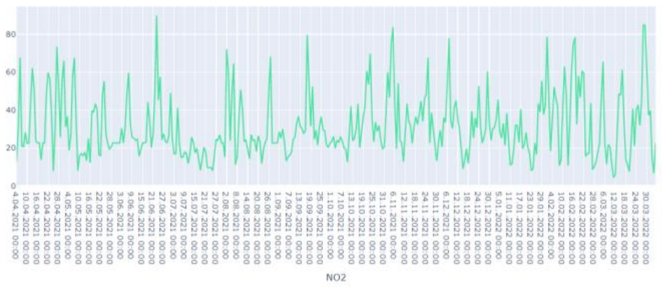


Fig. 10. Line Scatter for NO2

Figure 11 shows a balanced distribution in annual values. The highest value belongs to November.

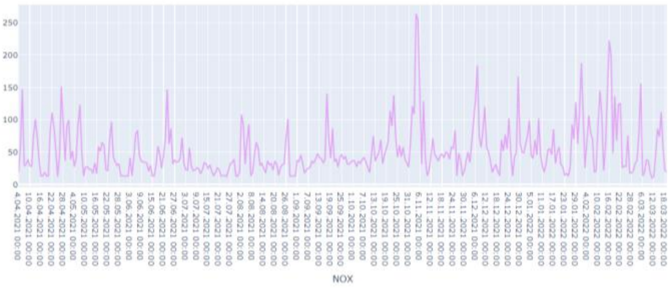


Fig. 11. Line Scatter for NOX

The annual data for Figure 12 are shown below. Annually the values are mostly high.

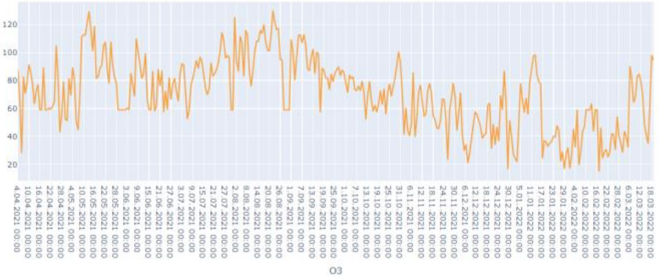


Fig. 12. Line Scatter for O3

The coloring process used for the bar plot is important in terms of both understanding the value of the data and seeing the peak values. Again, the images created on the same data set are shown in Figures 13-17 respectively. Figure 13 shows the distribution of annual PM10 values. High and low values are clearly visible and its coloring helps a lot.

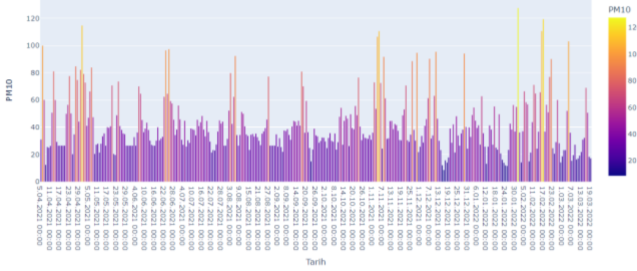


Fig. 13. Bar Plot for PM10

An uneven distribution is seen in Figure 14. High and low values are clearly visible thanks to both colors and the use of bar plots.

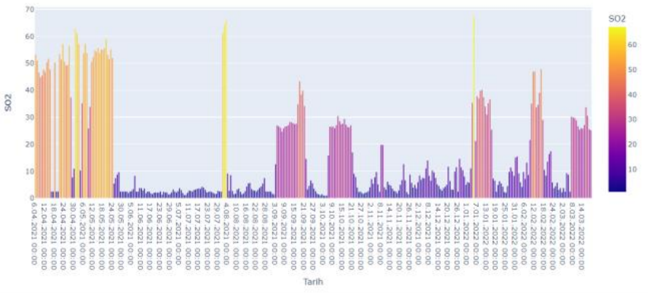


Fig. 14. Bar Plot for SO2

Annual NO2 values are shown in Figure 15. It is seen that there is a balanced distribution.

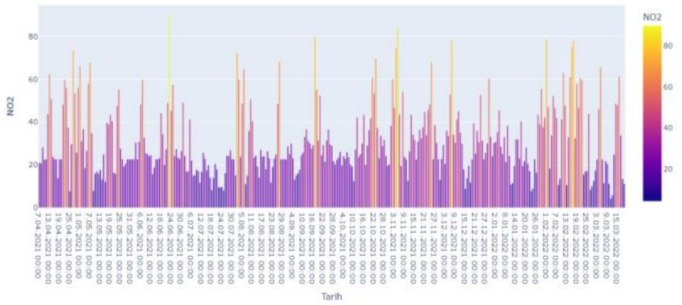


Fig. 15. Bar Plot for NO2

Figure 16 shows NOX values. High values are clearly visible. In general, it is seen to be at normal annual levels. The highest value belongs to November.

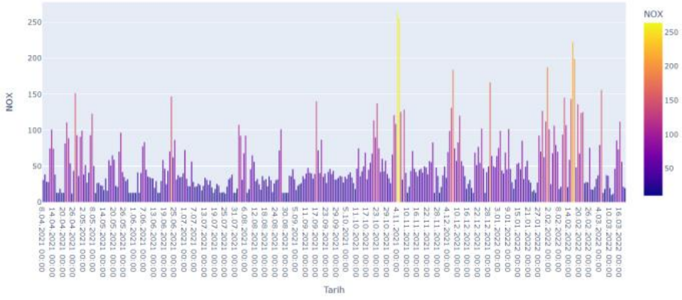


Fig. 16. Bar Plot for NOX

In Figure 17, it is seen that the annual O3 values are generally high.

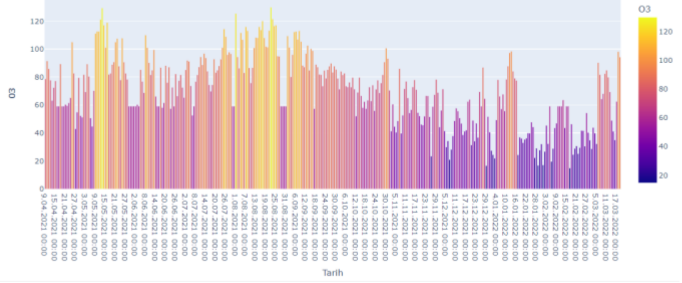


Fig. 17. Bar Plot for O3

Point scatter helped to show the distribution of air pollutants more clearly than other graphs. However, the line scatter allowed us to see the peak values more clearly. Like the point scatter plot, the bar plot allowed us to see the distribution more clearly as a result of coloring. In addition, the peaks can be easily seen in the bar plot.

V. Conclusion

In this study, trends in measured air pollution concentrations are visualized to make sense of large amounts of air data. One-year air pollutants data provided by the continuous monitoring center of the Turkish Ministry of Environment, Urbanization and Climate Change were obtained, and missing data were completed with the mean algorithm. Then, using data visualization methods such as point scatter plot, line scatter plot and bar plot, it was ensured that large amounts of data on environmental parameters were understandable by different stakeholders. Here, the excess and confused data has become more understandable with visualization. In future studies, it is planned to work on longer-term data, to use related data, to use different data visualizations suitable for these data types, to conduct a study that reveals the values of pollutants and what environmental effects are. Thus, researchers will be shown which visualization methods will have more appropriate use on which data types. Although we can see the distribution and minimum-maximum values with these three visualization methods used in our study, we could not get much detail about the data. This is a limiting feature of our study.

References

[1] B. Li, Z. Qiu, J. Zheng. **“**Impacts of noise barriers on near-viaduct air quality in a city: a case study in Xi’an”. Build. Environ., 196 (2021), Article 107751, 10.1016/j.buildenv.2021.107751

[2] K.F. Lu, H.D. He, H.W. Wang, X.B. Li, Z.R. Peng Characterizing temporal and vertical distribution patterns of traffic-emitted pollutants near an elevated expressway in urban residential areas Build. Environ., 172 (2020), Article 106678, 10.1016/j.buildenv.2020.106678

[3] H.D. He, H.O. Gao Particulate matter exposure at a densely populated urban traffic intersection and crosswalk Environ. Pollut., 268 (2021), Article 115931, 10.1016/j.envpol.2020.115931

[4] A. Lak, M. Ramezani, R. Aghamolae Reviving the lost spaces under urban highways and bridges: an empirical study J. Place Manag. Dev., 12 (2019), pp. 469-484, 10.1108/JPMD-12-2018-0101

[5] A. Sharma, D.D. Massey, A. Taneja A study of horizontal distribution pattern of particulate and gaseous pollutants based on ambient monitoring near a busy highway Urban Clim., 24 (2018), pp. 643-656, 10.1016/j.uclim.2017.08.003

[6] K.F. Lu, H.D. He, H.W. Wang, X.B. Li, Z.R. Peng Characterizing temporal and vertical distribution patterns of traffic-emitted pollutants near an elevated expressway in urban residential areas Build. Environ., 172 (2020), 10.1016/j.buildenv.2020.106678

[7] B. Das, Ö. O. Dursun, and S. Toraman, “Prediction of air pollutants for air quality using deep learning methods in a metropolitan city,” *Urban Climate*, (2022), vol. 46, p. 101291, , doi: 10.1016/j.uclim.2022.101291.

[8] C. Wu, H. He, R. Song, Z. Peng Prediction of air pollutants on roadside of the elevated roads with combination of pollutants periodicity and deep learning method Build. Environ., 207 (2022), Article 107436, 10.1016/j.buildenv.2021.108436

[9] G. Kurnaz, A.S. Demir Prediction of SO_2 and PM_{10} air pollutants using a deep learning-based recurrent neural network: case of industrial city Sakarya Urban Clim 41 (2021), Article 101051, 10.1016/j.uclim.2021.101051

[10] [10] P. Perez, C. Menares, C. Ramírez $PM_{2.5}$ forecasting in Coyhaique, the most polluted city in the Americas Urban Clim., 32 (2020), p. 100608, 10.1016/j.uclim.2020.100608

[11] A. Aggarwal, D. Toshniwal A hybrid deep learning framework for urban air quality forecasting**,** J. Clean. Prod., 329 (2021), Article 129660, 10.1016/j.jclepro.2021.129660

[12] M.Sülü, R. Daş, “Graph visualization of cyber threat intelligence data for analysis of cyber attacks”, Balkan Journal of Electrical and Computer Engineering (BAJECE), (2022),10(3), 300-306.

[13] M.Sülü, R. Daş, "QR Algoritması Kullanarak Spektral Çizge Bölümleme", Fırat Üniversitesi, Fen Bilimleri Dergisi, (2022), vol. 34, no. 2, pp. 207-218.

[14] C. Bachechi, L. Po, and F. Rollo, “Big Data Analytics and Visualization in Traffic Monitoring,” *Big Data Research*, (2022), vol. 27, p. 100292, doi: 10.1016/j.bdr.2021.100292.

[15] C. von Brömssen, S. Betnér, J. Fölster, and K. Eklöf, “A toolbox for visualizing trends in large-scale environmental data,” *Environmental Modelling & Software*, 2021, vol. 136, p. 104949, doi: 10.1016/j.envsoft.2020.104949.

[16] W. Huang, X. Xu, M. Hu, and W. Huang, “A license plate recognition data to estimate and visualise the restriction policy for diesel vehicles on urban air quality: A case study of Shenzhen,” *Journal of Cleaner Production*, 2022, vol. 338, p. 130401, doi: 10.1016/j.jclepro.2022.130401.

[17] G. Carro, O. Schalm, W. Jacobs, and S. Demeyer, “Exploring actionable visualizations for environmental data: Air quality assessment of two Belgian locations,” *Environmental Modelling & Software*, 2022, vol. 147, p. 105230, doi: 10.1016/j.envsoft.2021.105230.

[18] D. Pérez-Campuzano, L. Rubio Andrada, P. Morcillo Ortega, and A. López-Lázaro, “Visualizing the historical COVID-19 shock in the US airline industry: A Data Mining approach for dynamic market surveillance,” *Journal of Air Transport Management*, 2022, vol. 101, p. 102194, doi: 10.1016/j.jairtraman.2022.102194.

[19] K. R. Prasad, G. R. Kamatam, M. B. Myneni, and N. R. Reddy, “A novel data visualization method for the effective assessment of cluster tendency through the dark blocks image pattern analysis,” *Microprocessors and Microsystems*, 2022, vol.93, 104625, doi: 10.1016/j.micpro.2022.104625.

[20] A. Eldawy, M. Mokbel, A. Alharthi, A. Azaidy, K. Tarek, S. Ghani SHAHED: a MapReduce-based system for querying and visualizing spatio-temporal satellite data IEEE 31st International Conference on Data Engineering (2015), pp. 1585-1596, 10.1109/ICDE.2015.7113427

[21] G.Van Snickt, S. Legrand, J. Caen, F. Vanmeert, M. Alfeld, K. Jansses Chemical imaging of stained-glass windows by means of macro X-ray fluorescence (MA-XRF) scanning.Microchem. J. (2016), pp. 615-622

[22] A. Syed, N. Gupta, G. Nayak, R. Lenka Big Data Visualization: Tools and Challenges IEEE 2nd Int. Conference on Contemporary Computing and Informatics (2016), pp. 656-660, doi: 10.1109/IC3I.2016.7918044

[23] G. Carro, O. Schalm, W. Jacobs, and S. Demeyer, “Exploring actionable visualizations for environmental data: Air quality assessment of two Belgian locations,” *Environmental Modelling & Software*, 2022, vol. 147, p. 105230, doi: 10.1016/j.envsoft.2021.105230.

[24] SIM Air Quality- Station Data Download Continuous Monitoring Center. https://sim.csb.gov.tr/STN/STN_Report/StationDataDownloadNew (2020) (accessed 17 June 2022)

BIOGRAPHIES

DAMLA MENGÜŞ is currently studying in his Bachelor's degree in the Department of Software Engineering, Technology Faculty, at the Firat University. She works at the Department of Computer Engineering, Marmara University. Her current research areas include data visualization, machine learning, data science, and artificial intelligence.

BİHTER DAŞ graduated B.S. and M.S. degrees from the Department of Computer Science at the Firat University in 2004 and 2007 respectively. Then she received Ph.D. degree at the Department of Software Engineering at the same university in 2018. She also worked between September 2017 and June 2018 as a visiting scholar at the Department of Computing Science at the University of Alberta, Edmonton, Canada. Her current research areas include data science, big data, data analytics, bioinformatic, digital signal processing, genome data analysis.

Reflection Coefficient Calculation of a Structure Including a Porous Silicon Layer with Transfer Matrix Method and FDTD

Caglar Duman

Abstract—**Porous silicon is an important material for a variety of application area such as anti-reflective coating for solar cells. Today, solar cell market is mostly dominated by silicon based solar cells. Porous silicon thin films are easy to fabricate and it is compatible with silicon technology. Designing porous silicon anti-reflective coating layers is a critical issue to enhance silicon based solar cell performance. There are several methods to calculate reflection coefficient of porous silicon thin layers. In this study, transfer matrix method and finite-difference time-domain method are used to calculate reflection coefficient of porous silicon thin layers. Because finite-difference time-domain method gives more accurate results, the results obtained with finite-difference time-domain method are used to control the results obtained with transfer matrix method. In transfer matrix method, refractive indices of the porous silicon layers are calculated with Bruggeman effective medium approximation. A slab consists of 20 nm thick porous silicon layer on a 30 nm thick silicon layer free standing in the air is considered for the simulations. Porosity of the porous silicon layer is taken as 30%, 40% and 50%. Also, the porous silicon layer is considered as consisted of random placed pores with randomly changing diameters between 12 and 18 nm. The simulation results show that increasing the porosity and the pore diameters cause more divergence of transfer matrix method and finite-difference time-domain method results. Transfer matrix method results are more reliable for longer wavelengths because the porous silicon begins to resemble a homogeneous medium. In this study, it is aimed to investigate validity limits of transfer matrix method by comparing finite-difference time-domain method results. In the literature, there are several numerical and experimental studies investigating reflection coefficient of porous silicon. But best of our knowledge, there is no study investigating dependence of reflection coefficient on both the porosity and the pore sizes of porous silicon and validity limitation of transfer matrix method in the literature.**

Index Terms— **Anti-Reflective Coating, Solar Cell, Reflection Coefficient, TMM, FDTD.**

I. Introduction

ABUNDANCE OF silicon in the nature and its compatibility with today's electronic technology make it useful for a variety of applications. Porous silicon (PSi) is discovered in 1956 at Bell Labs, and it has great attention in 1990 by the scientific community with Leigh Canham's study. In following years, papers about its potential applications in microelectronics, optoelectronic devices, chemical and biological sensing are published. PSi is a sponge-like form of the silicon [1-3]. PSi is used for many applications in solar cells, fuel cells, biology, nanoenergetics, microelectromechanical systems, sensors, and photonic crystals [4]. An important feature of PSi is photoluminescence property. This property is because of excitonic recombination quantum confined in silicon nanocrystals. Although, its chemical instability, slow speed operation, disordered nature and fabrications problems are limits PSi usage for the photoluminescence applications, there are ongoing studies to overcome these problems [5]. Multilayer PSi structures are important for various areas such as optoelectronic and sensing applications. Some fabrication errors can cause significant changes optical properties of multilayer PSi structures. In a pioneer study investigating multilayer PSi structures with numerical and experimental methods, it is stated that with some basic precautions in multilayer PSi structure fabrication their optical performances can be increased [6]. In silicon thin film solar cells, layer or layers consist of PSi can be used as anti-reflective coating (ARC) which very necessary to enhance light trapping mechanism of the solar cell [2].

In [7], mathematical modeling of anti-reflective subwavelength structures is reviewed. The methods include effective medium theory (EMT), finite-difference time-domain (FDTD), transfer matrix method (TMM), the Fourier modal method (FMM)/rigorous coupled-wave analysis (RCWA) and the finite element method (FEM). All methods predict the broadband reflection of tapered nanostructures with periods smaller than the wavelengths of light of interest and lengths that are at least a large portion of the wavelengths. In [8], reflection spectra are obtained for designed nanostructure geometries on amorphous silicon thin-film solar cells, using a Ray Tracing modelling approach. This coating reduced reflectance in the wavelength of 300–800 nm range by an average of 2.665% and 11.36% at 0◦ and 80◦ incident light, respectively. A reflectance reduction of 19.192% is obtained for wavelength of 300 nm and 80◦ incident light.

To design ARCs consisted of PSi layer or layers, refractive index of the PSi is to be known. Bruggeman effective medium

ÇAĞLAR DUMAN, is with Department of Electrical and Electronic Engineering University of Erzurum Technical University, Erzurum, Turkiye, (e-mail: caglarduman@erzurum.edu.tr).

https://orcid.org/0000-0002-1845-8605

Manuscript received Jul 18, 2022; accepted Oct 25, 2022.
DOI: 10.17694/bajece.1144847

approximation (EMA) is an effective method to calculate refractive index of the PSi but it is reliable if wavelength of incident light is much larger than the pore sizes of the PSi. If refractive index of the PSi is known by using transfer matrix method (TMM), reflection coefficient of the PSi including structure can be calculated. In the literature several studies, using TMM to analysis structures including PSi [9-11]. FDTD method also can be used to the reflection coefficient [12, 13]. While using FDTD method, knowing refractive index of the PSi is not necessarily so FDTD gives more accurate reflection coefficient results. In [14], FDTD calculation and reflectance measurement are performed for PSi with porosity equal to 60% and it is found that FDTD calculations agree with the measured reflectance. Best of our knowledge, there is no study investigating dependence of reflection coefficient both on the porosity and the pore sizes of porous silicon and validity limitation of transfer matrix method in the literature.

In the study, a two layered structure, consists of 20 nm thick PSi layer on a 30 nm thick silicon layer free standing in the air is considered. For different porosities and pore sizes, TMM and FDTD simulations are performed to observe limitation of TMM method. In the 2nd and 3rd sections of this manuscript, information about pore structure of PSi thin films and calculation of refractive index of PSi are presented, respectively. In 4th and 5th sections of the manuscript, reflection coefficient calculation with TMM and FDTD are explained. In the 6th section, the model used in simulations and obtained results are given. Finally, in the Conclusions section, the results are summarized and argued.

II. Porous Silicon Fabrication

Even though, there are more than 20 methods to fabricate PSi structures, electrochemical etching is the most preferred method to fabricate PSi structures because it is simple and economical. This method is performed in hydrofluoric (HF) solution but just dipping the silicon in HF is not formed PSi. A current flow between two electrodes, which is the silicon at the anode and platinum at the cathode is needed [3, 15]. In electrochemical etching, current density, HF concentration, temperature, etching time, type of silicon and dopant concentration are the parameters effecting formation of the pores [16].

There are ongoing debates about exact pore formation mechanism [4]. But there are studies in the literature showing selective pore direction is <100>. For (001) oriented silicon substrates, there is only one <100> direction which is perpendicular to the surface thus the pores form in well-defined columnar structure [17]. From TEM images of such PSi layers it is seen that it is possible to obtain pores with smooth walls and homogeneous thicknesses without inter connections between the pores [3]. In this study, the pores are considered as randomly distributed cylindrical structures perpendicular to the surface with smooth walls. Because of their random distribution they can overlap with each other and form a more complex pattern.

III. Refractive Index of Porous Silicon

Porosity of a PSi wafer is show air volume in the structure and determined as in Equation 1 [1].

$$P = \frac{m_1 - m_2}{m_1 - m_3} \qquad (1),$$

where P is the porosity, m_1, m_2 and m_3 are the results from weight measurements before the anodic reaction, after the anodic reaction and finally after dissolution of the porous material in a molar acidic aqueous solution, respectively. Pores in a PSi are randomly arranged and the PSi can be group according to pore sizes. If the pore sizes are less than 4 nm it named as nanoporous silicon and if the pore sizes are in the range of 4 – 50 nm it named as mesoporous silicon [14]. By using Bruggeman EMA, refractive index of the PSi can be determined. EMA approximation is acceptable while wavelength of the incidence light is much larger than the pore sizes. Under this condition, the incident light does not distinguish the silicon and the void (air) and the PSi can be considered as a homogeneous medium. From Bruggeman EMA, the dielectric constant of a PSi layer can be determined as in Equation 2 [1, 18].

$$(1-P)\left(\frac{\varepsilon_{r,Si} - \varepsilon_{r,eff}}{\varepsilon_{r,Si} + (d-1)\varepsilon_{r,eff}}\right) + P\left(\frac{\varepsilon_{r,air} - \varepsilon_{r,eff}}{\varepsilon_{r,air} + (d-1)\varepsilon_{r,eff}}\right) = 0 \ (2),$$

where $\varepsilon_{r,Si}$, $\varepsilon_{r,eff}$, and $\varepsilon_{r,air}$ are the dielectric constants of silicon, PSi and air, respectively. In Equation 2, d is the system dimensionality and it is 3 for the nanoporous silicon and 2 for the mesoporous silicon [18]. The calculated effective dielectric constant ($\varepsilon_{r,eff}$) can be used to determine complex refractive index of the PSi ($\tilde{n}$) by using equality of $\varepsilon_{r,eff} = \tilde{n}^2$. By using Equation 2, refractive index variation with porosity of a nanoporous silicon and mesoporous silicon is obtained as shown in Figure 1.

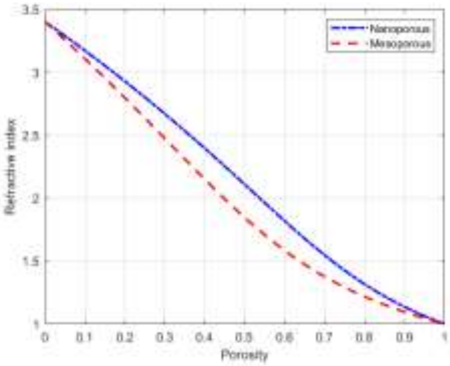


Figure 1. Refractive index variation with porosity for nanoporous silicon and mesoporous silicon.

It is seen from Figure 1, refractive index of the PSi varies between refractive index of silicon which is taken as 3.4 and refractive index of the air and with it is decrease with increasing of the porosity, as expected.

IV. Reflection Coefficient Calculation with TMM

After calculating the refractive index, reflection coefficient of the layers composed of a PSi layer and a silicon layer can be calculated. For the calculation, transfer matrix of the structure is generated as in Equation 3 [19].

$$T_l = \begin{bmatrix} (1+p_l)e^{-i\tilde{n}_l d_l \omega/c} & (1-p_l)e^{-i\tilde{n}_l d_l \omega/c} \\ (1-p_l)e^{+i\tilde{n}_l d_l \omega/c} & (1+p_l)e^{+i\tilde{n}_l d_l \omega/c} \end{bmatrix} \qquad (3).$$

In Equation 3, $\tilde{n}_l = n_l \cos\theta$ where n_l is refractive index of the l^{th} layer and θ is the angle of incidence. Also, d_l, ω and c show thickness of the l^{th} layer, radial frequency of the incident wave and speed of light, respectively. Finally, p_l is expressed as in Equation (4).

$$p_l(\omega) = \begin{cases} \dfrac{\tilde{n}_{l-1}}{\tilde{n}_l} & \text{for TE waves} \\[2ex] \dfrac{\tilde{n}_{l-1}\, n_l^2}{\tilde{n}_l\, n_{l-1}^2} & \text{for TM waves} \end{cases} \qquad (4).$$

The overall transfer matrix is obtained by sequentially multiplying the transfer matrices. The obtained transfer matrix can be considered as in Equation (5).

$$T = \begin{bmatrix} a & b \\ b^* & a^* \end{bmatrix} \qquad (5),$$

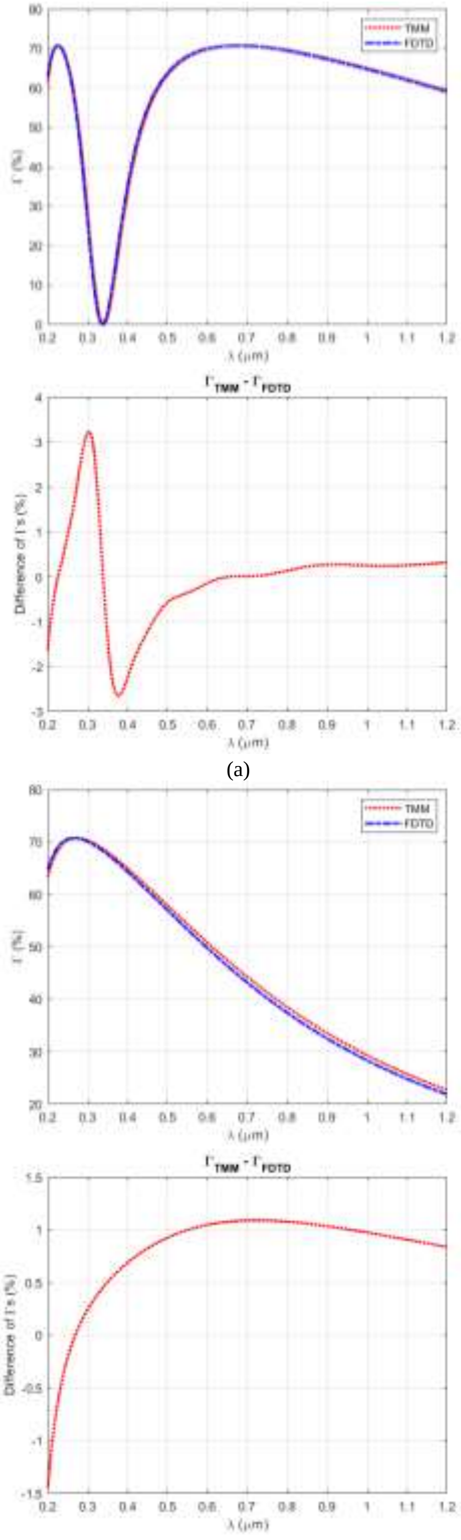
where * is denotes the complex conjugate. Reflection coefficient of the structure is given as in Equation (6).

$$\Gamma = -\frac{b^*}{a^*} \qquad (6).$$

V. Reflection Coefficient Calculation with FDTD Method

FDTD method is first described by Yee in 1966. It solves time dependent Maxwell's equations by approximating time and space as finite differences. FDTD method can be used for solving electromagnetic problems involving arbitrary geometries with no approximations other than curved structures modelled with a stair step approximation which cause errors [20, 21]. Also, FDTD method needs great computation power and time but it can simulate electromagnetic waves interaction with actual geometries more accurately [8].

For calculating reflection coefficient of a slab, a technique that integrating sources far from the problem zone into the FDTD method can be used. This method is known as scattered field formulation. Far-zone sources generate the incident fields outside the FDTD problem space and so there are no scatterers.



(a)

(b)

Figure 2. Reflection coefficients and their differences for porosity of a) 0% and b) 100%.

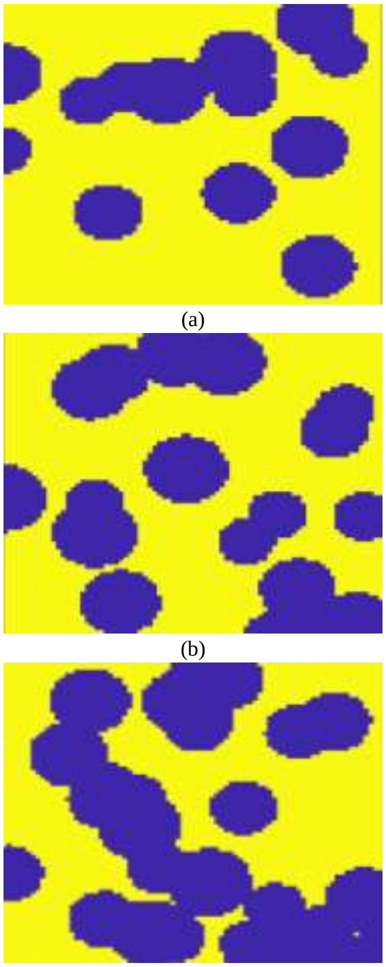
The field in the FDTD problem space is called total field and it is sum of the incident field and the scattering field. By rearranging FDTD equations accordingly, the scattered field and the incident field can be calculated while the simulation is running. The reflected field from the slab is the scattered field, and the ratio of the reflected field to incident field is the reflection coefficient as shown in Equation 7 [19].

$$|\Gamma| = \frac{\left|\vec{E}_{scat}\right|}{\left|\vec{E}_{inc}\right|} \qquad (7).$$

For a slab including PSi layer, the scattered field should be sampled at a surface below the slab, and the incident field should be sampled at the same surface. In this study, 3D FDTD calculation with CPML boundaries are applied and detailed explanation of the FDTD calculations is as in [22].

VI. Model and Simulation Results

In the study, a slab consists of 20 nm thick PSi layer on a 30 nm thick silicon layer free standing in the air is considered. Porosity of the PSi layer is taken as 30%, 40% and 50% and the refractive indices for these porosities is calculated with Bruggeman EMA. For FDTD simulation a slab with same thickness is considered. It is assumed that the slab is lies in the xy plane by penetrating into the CPML boundaries for 16 cells long in positive and negative x and y directions. 10 cells long air buffer and 16 cells long CPML layer are placed in positive and negative z boundaries. The width and length of the slab are 80 nm.

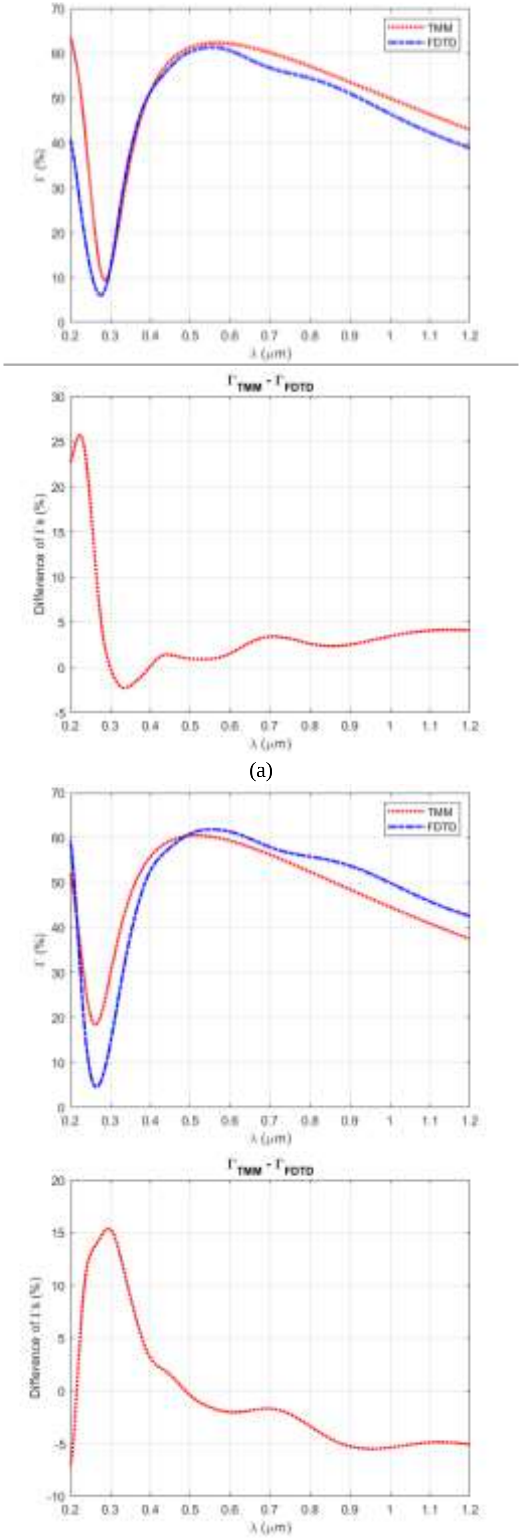


(a)

(b)

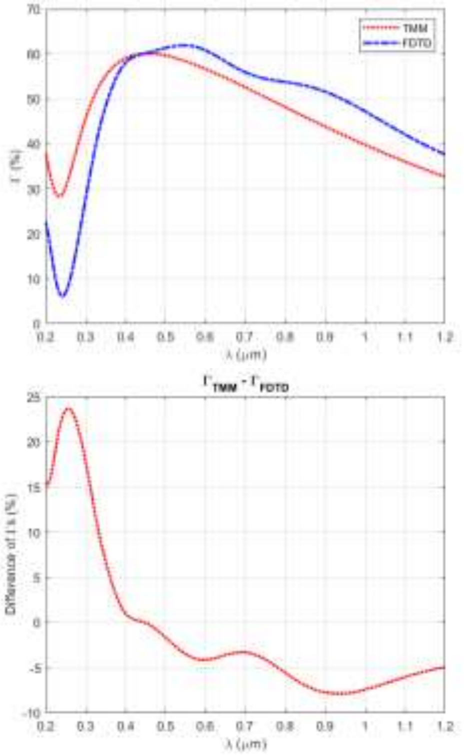
(c)

Figure 3. Obtained PSi layers for porosities of (a) 30%, (b) 40% and (c) 50%.



(a)

(b)



(c)

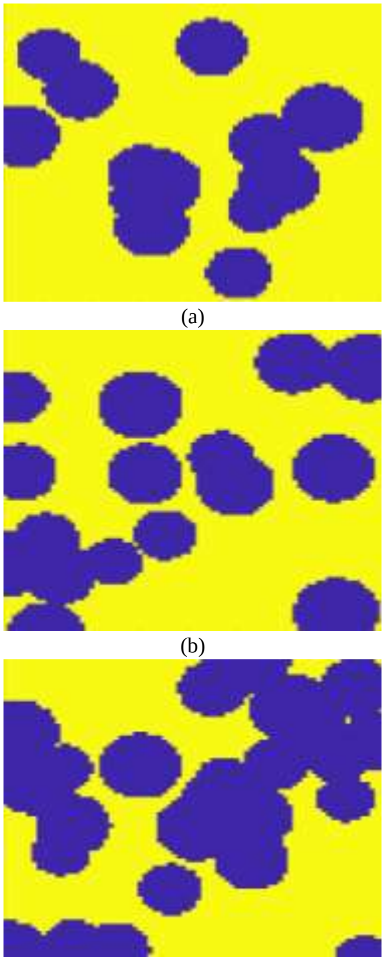
Figure 4. Obtained reflection coefficient and their differences for PSi layers with porosities of (a) 30%, (b) 40% and (c) 50%.

The incident light is x-polarized derivative gaussian shaped planar wave travelling in positive z direction. The scattered and incident fields are sampled at a surface 5 cells below the FDTD problem space. To test FDTD program, the reflection coefficient with FDTD and TMM are calculated for 0% porosity and %100 porosity. Porosity of 0% means no void in the PSi layer and porosity of 100% means no PSi layer. The results are shown in Figure 2.

It is seen from Figure 2-a and -b, the results obtained with TMM and FDTD methods well agrees but there are small differences because of the numerical errors which occur in the FDTD method. By reducing the spatial step size and the simulation time the results can be improved. In FDTD simulations, the PSi layer is created by random placed pores with randomly changing diameters between 12 and 18 nm and it is considered that the pores are perfectly continuous until the end of the PSi layer. The number of pores and their sizes are adjusted so as to reach the desired porosity. Obtained PSi layers for porosities of 30%, 40% and 50% are as shown in Figure 3.

In Figure 3, silicon portion of the PSi is shown with yellow and air portion of the PSi is shown with blue. Obtained reflection coefficient for PSi layers with porosities of 30%, 40% and 50% are shown in Figure 4.

It is seen from the figure, with increasing porosity the differences of the obtained reflection coefficients are also increase. Moreover, the results are more similar at high wavelengths and the differences decrease with increasing wavelength. The results obtained by using TMM is become more accurate at higher wavelengths because EMA approximation is acceptable while wavelength of the incidence light is much larger than the pore sizes. Under this condition, the incident light does not distinguish the silicon and the void (air) and the PSi can be considered as a homogeneous medium [1]. To observe effect of pore distribution on FDTD simulations, second set of PSi layers with same porosities are considered. Distribution of the pores are different from previous PSi layers because of their random nature. Obtained PSi layers are shown in Figure 5.

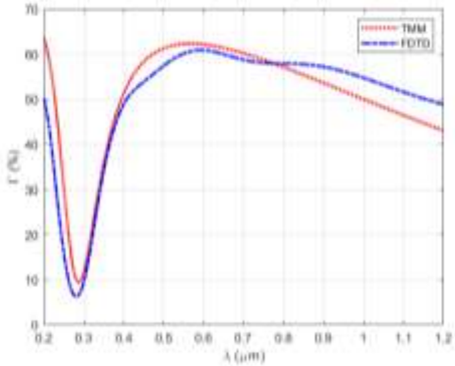

(a)

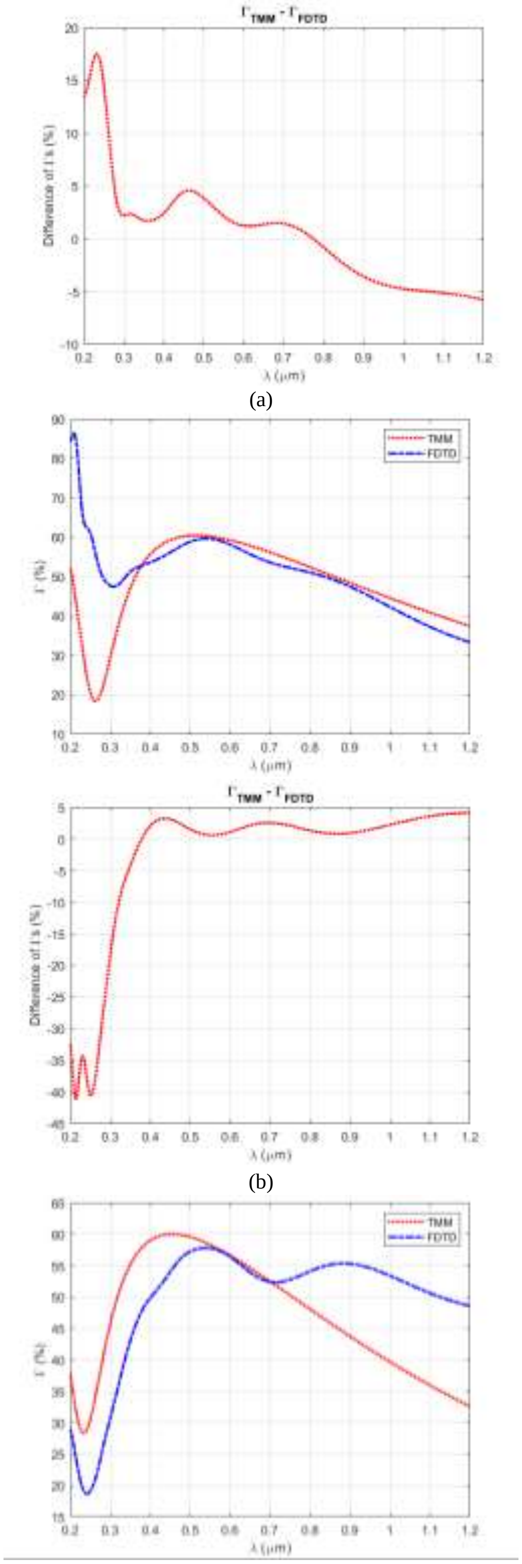
(b)

(c)

Figure 5. Second set of PSi layers for porosities of (a) 30%, (b) 40% and (c) 50%.

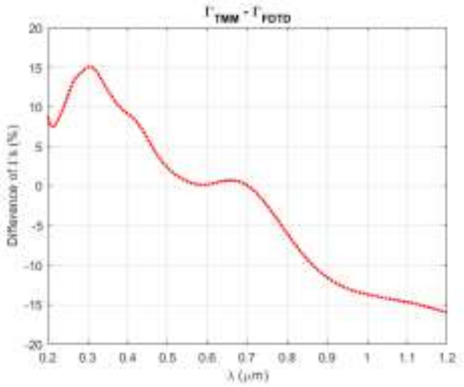
Obtained reflection coefficient of PSi layers with porosities of 30%, 40% and 50% are as in Figure 6.





(a)

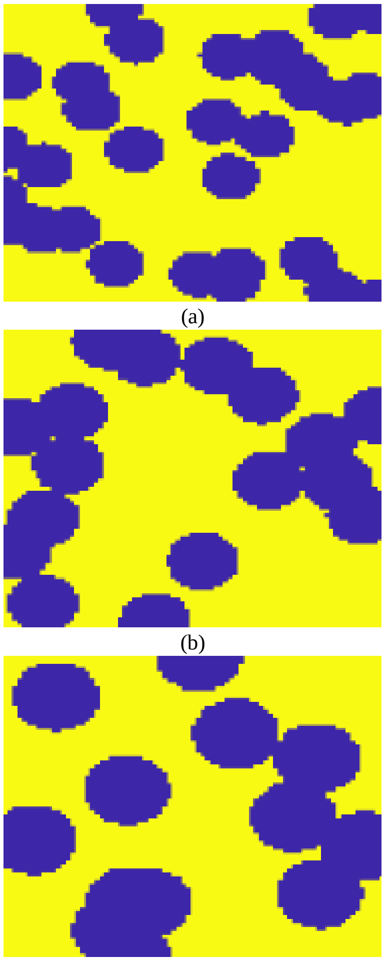
(b)



(c)

Figure 6. Obtained reflection coefficient of PSi layers and their differences with porosities of (a) 30%, (b) 40% and (c) 50%.

Even though results shown in Figure 6 are different from results shown in Figure 4, they show similar variations. The pore sizes are fixed 12 nm, 15 nm and 18 nm for 40% porosity and effect of pore size is evaluated. Obtained PSi layers can be seen in Figure 7.



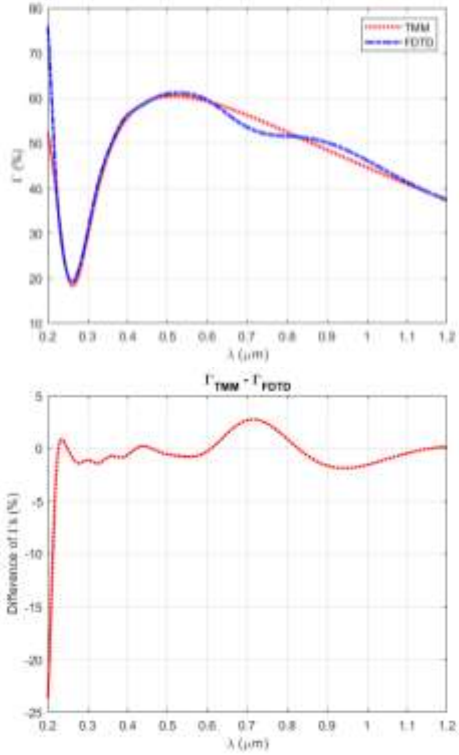
(a)

(b)

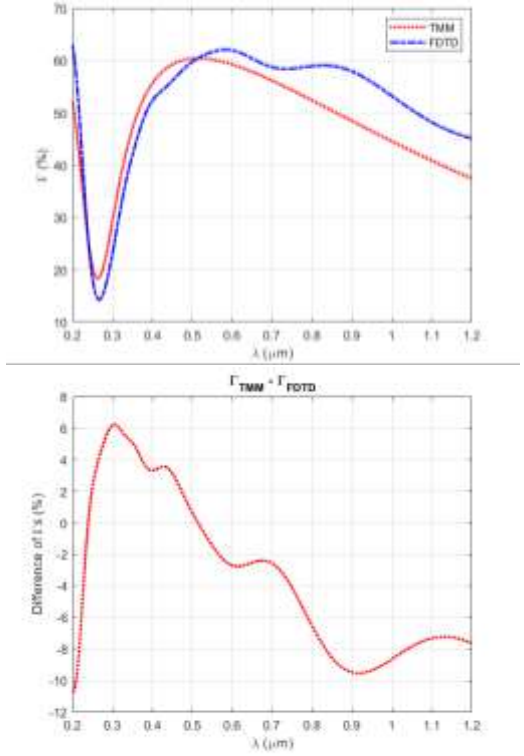
(c)

Figure 7. Obtained PSi layers with 40% porosity for pore diameters are (a) 12 nm (b) 15 nm and (c) 18 nm.

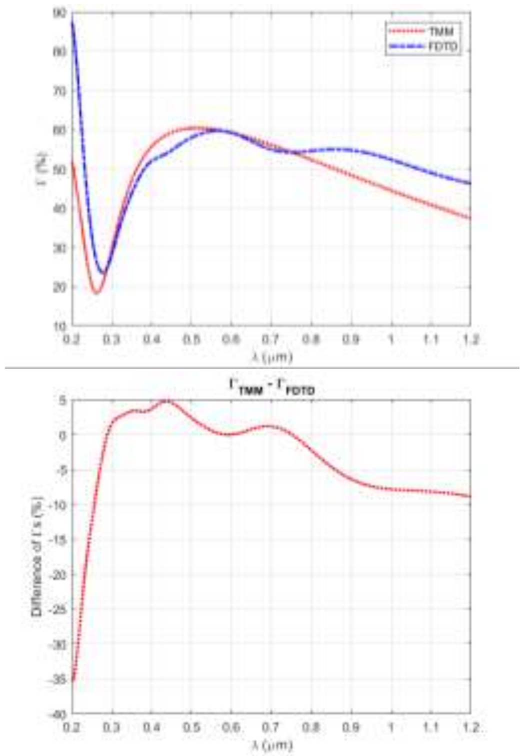
Obtained reflection coefficient of PSi layers with 40% porosity for pore diameters are 12 nm, 15 nm and 18 nm are as in Figure 8.



(a)



(b)



(c)

Figure 8. Obtained reflection coefficient of PSi layers and their differences with 40% porosity for pore diameters are (a) 12 nm (b) 15 nm and (c) 18 nm.

It is observed from Figure 8 that more accurate results obtain by using TMM for much longer wavelengths. This can be explained by the results obtained by using TMM becoming more accurate at higher wavelengths.

VII. Conclusions

In this study, reflection coefficient of PSi layer on a thin silicon layer is calculated with TMM and FDTD. Effect of pore size and pore distribution on the reflection coefficient are also studied. FDTD method solves electromagnetic problems with arbitrary geometries and simulate electromagnetic waves interaction with PSi more accurately. The errors in FDTD method are caused by numerical errors and stair step approximation. But in TMM, refractive index of the PSi is to be calculated and the PSi is considered as a homogeneous medium with a constant refractive index. Thus, by comparing results obtained with TMM and FDTD methods, accuracy of TMM method can be evaluated for structures including a PSi layer or layers.

It is seen from the simulation results that TMM are more reliable for small porosity values. Increasing the porosity causes more divergence of TMM and FDTD results. TMM results are seem more reliable for longer wavelengths because the wavelength is much higher than the pore size. With this condition, the PSi begins to resemble a homogeneous medium. In addition, since the pore size and pore distribution of PSi are random, the reflection coefficient vary depending on the PSi layer on which the calculations are made. However, since the refractive index of the medium is considered as constant in TMM, this can only be observed from the FDTD results.

The analyzes carried out in this study show that PSi ARC analysis made by using TMM will be reliable if wavelength of the incident light is much higher than the pore sizes. To obtain reliable results at lower wavelengths FDTD method should be preferred. The results are also important for practical applications. By using optical analysis during design phase of a practical application, expensive reworks can be avoided. At the design phase, if TMM with EMA would be used, porosity of the PSi should be low and pore sizes of the PSi should be smaller than interested wavelength range. Otherwise, methods such as FDTD should be used to obtain realistic results.

REFERENCES

[1] Basu, S. (Ed.). (2011). Crystalline Silicon: Properties and Uses. BoD–Books on Demand.

[2] Dubey, R. S., & Gautam, D. K. (2011). Porous silicon layers prepared by electrochemical etching for application in silicon thin film solar cells. Superlattices and Microstructures, 50(3), 269-276.

[3] Karbassian, F. (2018). Porous silicon. In Porosity-Process, Technologies and Applications. IntechOpen.

[4] Zhao, M., Balachandran, R., Allred, J., & Keswani, M. (2015). Synthesis of porous silicon through interfacial reactions and measurement of its electrochemical response using cyclic voltammetry. RSC advances, 5(96), 79157-79163.

[5] Bisi, O., Ossicini, S., & Pavesi, L. (2000). Porous silicon: a quantum sponge structure for silicon based optoelectronics. Surface Science Reports, 38(1-3), 1-126.

[6] Hasar, U. C., Özbek, İ. Y., & Karacalı, T. (2017). Optical Characterization of Porous Silicon Multilayers. Handbook of Porous Silicon, Editor: Canham Leigh, Springer, 1-12.

[7] Han, K., & Chang, C. H. (2014). Numerical modeling of sub-wavelength anti-reflective structures for solar module applications. Nanomaterials, 4(1), 87-128.

[8] Pickering, T., Shanks, K., & Sundaram, S. (2021). Modelling technique and analysis of porous anti-reflective coatings for reducing wide angle reflectance of thin-film solar cells. Journal of Optics, 23(2), 025901.

[9] Wilkins, M. M., Boucherif, A., Beal, R., Haysom, J. E., Wheeldon, J. F., Aimez, V., ... & Hinzer, K. (2013). Multijunction solar cell designs using silicon bottom subcell and porous silicon compliant membrane. IEEE Journal of Photovoltaics, 3(3), 1125-1131.

[10] Ariza-Flores, D., Pérez-Huerta, J. S., Kumar, Y., Encinas, A., & Agarwal, V. (2014). Design and optimization of antireflecting coatings from nanostructured porous silicon dielectric multilayers. Solar energy materials and solar cells, 123, 144-149.

[11] Jimenéz-Vivanco, M. R., García, G., Carrillo, J., Morales-Morales, F., Coyopol, A., Gracia, M., ... & Lugo, J. E. (2020). Porous Si-SiO2 UV Microcavities to modulate the responsivity of a broadband photodetector. Nanomaterials, 10(2), 222.

[12] Deinega, A. V., Konistyapina, I. V., Bogdanova, M. V., Valuev, I. A., Lozovik, Y. E., & Potapkin, B. V. (2010). Optimization of an anti-reflective layer of solar panels based on ab initio calculations. Russian Physics Journal, 52(11), 1128.

[13] Min-Dianey, K. A. A., Zhang, H. C., Brohi, A. A., Yu, H., & Xia, X. (2018). Optical spectra of composite silver-porous silicon (Ag-pSi) nanostructure based periodical lattice. Superlattices and Microstructures, 115, 168-176.

[14] Najar, A., Al-Jabr, A. A., Slimane, A. B., Alsunaidi, M. A., Ng, T. K., Ooi, B. S., ... & Anjum, D. H. (2013, April). Effective antireflection properties of porous silicon nanowires for photovoltaic applications. In 2013 Saudi International Electronics, Communications and Photonics Conference (pp. 1-4). IEEE.

[15] Burham, N., Hamzah, A. A., & Majlis, B. Y. (2017). Self-adjusting electrochemical etching technique for producing nanoporous silicon membrane. New Research on Silicon-Structure, Properties, Technology.

[16] Yaakob, S., Bakar, M. A., Ismail, J., Bakar, N. H. H. A., & Ibrahim, K. (2012). The formation and morphology of highly doped N-type porous silicon: effect of short etching time at high current density and evidence of simultaneous chemical and electrochemical dissolutions. J. Phys. Sci, 23(2), 17-31.

[17] Vázsonyi, É., Battistig, G., Horváth, Z. E., Fried, M., Kádár, G., Pászti, F., ... & Poortmans, J. (2000). Pore Propagation Directions in P+ Porous Silicon. Journal of Porous Materials, 7(1), 57-61.

[18] Khardani, M., Bouaïcha, M., & Bessaïs, B. (2007). Bruggeman effective medium approach for modelling optical properties of porous silicon: comparison with experiment. Physica Status Solidi c, 4(6), 1986-1990.

[19] Birge, J. R., & Kärtner, F. X. (2006). Efficient analytic computation of dispersion from multilayer structures. Applied Optics, 45(7), 1478-1483.

[20] Pérez, E. X. (2008). Design, fabrication and characterization of porous silicon multilayer optical devices. Universitat Rovira i Virgili.

[21] Hossain, M. F., & Noushin, T. (2016, December). Sensitivity enhancement of porous silicon waveguide sensor using graphene by FDTD with Lumerical software. In 2016 2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE) (pp. 1-4). IEEE.

[22] Elsherbeni, A. Z., & Demir, V. (2009). The finite-difference time-domain method for electromagnetics with MATLAB simulations. Raleigh, NC: SciTech Pub.

BIOGRAPHIES

ÇAĞLAR DUMAN was born Erzurum, in 1981. He received the B.S. and M.S. degrees in electronic engineering from the Niğde Ömer Halisdemir University in 2005 and Ataturk University in 2008, respectively, and the Ph.D. degree in electronic engineering from Ataturk University, Erzurum in 2014. Since 2014, he has been Assistant Professor in Department of Electrical and Electronic Engineering University of Erzurum Technical University. His research interests include lasers and photovoltaics.

Hybrid Convolutional Neural Network Method for Robust Brain Stroke Diagnosis and Segmentation

Sercan Yalcin

Abstract— **Artificial intelligence with deep learning methods has been employed by a majority of researchers in medical image classification and segmentation applications for many years. In this study, a hybrid convolutional neural network (CNN) model has been proposed for diagnosing brain stroke from the dataset consisting of computed tomography (CT) brain images. The model inspired by C-Net consists of multiple concatenation layers of the networks, and prevents the concatenation of convolutional feature maps to evince the mapping process. The structures of the convolutional index and residual shortcuts of the INet model are also integrated into the proposed CNN model. In the output layer of the model, it is split into two classes as whether there is a stroke or not in a brain image, and then the region of the stroke in the image is segmented. Tremendous analyzes have been conducted in terms of many benchmarks using Python programming. The proposed method shows better performance than some other current CNN-based methods by 99.54% accuracy and 99.1% Matthews correlation coefficient (MCC) in the diagnosis of brain stroke. The proposed method can alleviate the work of most medical staffs and facilitate the process of the patient's remedy.**

Index Terms— **Artificial intelligence, brain stroke diagnosis, convolutional neural networks, deep learning.**

I. Introduction

IT IS HOPED that humanity will use artificial intelligence-based medical technologies to create applications that help society by fusing social responsibility consciousness with technological understanding [1].

Brain stroke is one of the deadliest diseases in the world and rapid diagnosis is very important in the medical treatment process. [2]. Patients who exhibit certain symptoms may have had an ischemic or hemorrhagic stroke. When blood clots prevent or drastically restrict blood flow to the brain, an ischemic stroke results. After an ischemic stroke, patients may also develop stroke bleeding, which is a dangerous consequence.

SERCAN YALÇIN, is with Department of Computer Engineering University of Adiyaman University, Adiyaman, Turkey (e-mail: svancin@adiyaman.edu.tr).

https://orcid.org/0000-0003-1420-2490

Manuscript received June 11, 2022; accepted Aug 17, 2022.
DOI: 10.17694/bajece.1129233

While hemorrhage happens as a result of a stroke, blood traveling to other surrounding brain tissues, blood vessels bursting due to their rigidity, or both. A concussion, high blood pressure, bleeding problems, aneurysms, and arteriovenous malformation are the main causes of hemorrhagic brain stroke [3]. The care and outlook for stroke patients must be improved because it is well known that strokes are a severe health issue. Therefore, quick and accurate diagnostic techniques are required. The terms "brain imaging techniques" relate to magnetic resonance imaging (MRI) and computed tomography (CT). They give the doctors the patient consults a hint as to how to keep the patient under initial control. Additionally, there are a number of imaging methods for examining the brain, such as magnetoencephalography, functional magnetic resonance imaging, emission positron tomography, and X-ray and optical imaging. The most widely used imaging technique is the CT scan. This is mostly because patients may receive its images, which are less expensive than those from other imaging systems. The first step in providing patients with an appropriate diagnosis and course of treatment is the ability to predict brain stroke using CT imaging [4].

Convolutional neural networks (CNNs), which are deep learning techniques based on artificial intelligence, have made significant progress in the recognition of biomedical images. When working with medical images, CNNs are employed for semantic segmentation procedures where each pixel in the image is labeled by a neighboring object or region. Along with classification, picture segmentation is a crucial job that is used to increase the accuracy of diagnoses. The primary goal and task of the medical image, which goes through a pixel-level categorization procedure, is actually image segmentation [5]. A deep CNN-based method for the identification and classification of acute ischemic stroke was presented by Lo et al. [6] utilizing CT images. The imaging dataset for the is made up of 573 CT scans from 96 patients who had ischemic strokes and 96 healthy controls (681 images). Radiologists were able to diagnose acute ischemic stroke thanks to transfer learning, which was successful in establishing a specific scratch training approach for a specific scanner. On MRI pictures of patients, Tomitaa et al. [7] suggested a deep neural network approach for segmenting severely wounded brain lesions. A total of 239 pictures from a dataset of patients with persistent ischemic stroke were processed using the suggested scheme. A new zooming technique was used in performance analyses of 3D segmentation models with residual networks. Deep learning and CNN were suggested by Gaidhani et al. [5] as a technique for identifying brain stroke using an MRI. Brain stroke MRI pictures might be separated into normal and abnormal images using the suggested strategy. Semantic segmentation was also used to identify anomalous regions. LeNet and SegNet are two different CNN types that are utilized by the suggested methodology [8]. LeNet is a CNN deep architecture built on encoder-decoder technology. SegNet is a sophisticated and complete CNN model for semantic pixel-wise segmentation. For the purpose of segmenting images, the encoder-decoder CNN architecture known as UNet [9] was created. This architecture's major objective is to provide a shared negotiation network with successive layers that uses upsampling operators in place of pooling operators. In order to dynamically separate acute ischemic stroke lesions from multi-directional MRIs, Liu et al. [10] presented a new deep residual CNN. Any enhancements made by INet [12] are guaranteed when the original U-Net contains residual shortcuts known as ResUNet [11]. The degradation issue was improved by the Res-CNN using additional data from MRIs [12]. The definition of ResDenseUNet [13], a different model to compare with the suggested model with dense linkages known as DenseINet [12], is when residual shortcuts are added to the original DenseUNet [12]. Data augmentation and data aggregation techniques were employed to enhance the amount of training images prior to training the network model. On two acute ischemic stroke datasets, seven neural networks were trained, and the outcomes were thoroughly examined. CNN combined with random forests was used by Saragih et al. [14] to conduct ischemic stroke detection based on a patient's CT scan. In this approach, in the categorization of data based on feature extraction with CNN, the fully linked layer has been replaced by completely random forests. 10% of the data set was test data when the suggested procedure was applied. A new CNN design dubbed C-Net that combines various networks was proposed by Barzekar and Yu [15]. On the BreakHis and Osteosarcoma datasets, the C-Net was used for the categorization of histological picture. For both datasets, the C-Net model was effective, yielding no misclassifications. For the purpose of identifying movement-related brain MR artifacts, Oksuz [16] proposed dense CNNs and a residual U-Net architecture. A method based on MR physics was used to create artificial artifacts. A residual U-net network tuned using corrupted data helped to improve the observed artifacts. The architecture, which handles artifact detection and correction, produced higher-quality images and made it possible to segment brain strokes more precisely. A novel technique for cerebral vascular segmentation without the requirement for physical intervention was put out by Deshpande et al. [17]. In order to disclose vascular geometric aspects and categorize vascular anatomy, the scientists also provided a model by skeletonizing the binary segment map. They divided MR and CT angiograms using an active contour-based method. This method combines probabilistic grain-enhancing filtering with a Hessian framework. Additionally, the vessel centerlines and diameters have been calculated using this method in order to determine the geometrical characteristics of the vasculature. Dimension-fusion-UNet (D-UNet), a brain segmentation model suggested by Zhou et al. [18], mixes 2D and 3D convolution during the encoding phase. In comparison to 2D networks, the suggested model performs better in segmentation. Compared to 3D networks, the model requires substantially less computational effort. To lessen the data imbalance between positive and negative instances for the network's training, the scientists also developed a new loss function called Enhance Mixing Loss (EML).

The brain CT pictures have been examined in this paper to assess whether or not a stroke has occurred. Additionally, the section of the brain that the radiologist examined has been calculated after the brain strokes in the photographs were segmented. A deep learning algorithm based on CNN has been suggested. In the brain CT images of the dataset collected from the Ministry of Health of the Republic of Turkey, the suggested model recognizes and categorizes brain strokes. By using the segmentation method, the model can identify and forecast the stroke region. Other current CNN models, including ResNet50v2 [19], UNet [9], DeepLabV3 [20], ResUNet [11], DenseINet [12], ResDenseUNet [13], and C-Net [15], have also been used to examine performance for classifying and segmenting brain strokes in the same CT brain images. A variety of evaluations have been performed based on comparisons between the performance results that were produced.

The following are the paper's main contributions:

- To classify and segment the brain pictures, a deep hybrid model based on C-Net and INet is provided. The proposed model might be crucial for the quick identification and treatment of brain stroke.
- Preprocessing techniques have been used to improve the contrast of the image, making it easier to identify the stroke-affected area from the brain scans. Additionally, the proposed model successfully designs the layers of convolution, pooling, dropout, and fully connected, and it takes convolutional index and residual shortcuts into account.
- The Ministry of Health of the Republic of Turkey provided the study with an actual data set. The collection contains CT images in the Digital Imaging and Communications in Medicine (DICOM) format that are not stroke-related, ischemic, hemorrhagic, or overlay.
- According on experimental performance findings, the suggested model outperforms DenseINet, ResDenseUNet, and C-Net.
- Although this study has been inspired by our previous study [21] and the data set used was the same, it can be said that it is original in terms of the proposed method and an innovative approach is presented.

The rest of the paper organization is as follows: Section 2 explains the material and method. Experimental analyzes and results are presented in Section 3. Both classification and segmentation results are evaluated, and then compared to other recent studies. Finally, in Section 4, the study is concluded.

II. Material And Method

In this study, a new hybrid deep learning scheme is proposed for brain stroke diagnosis by taking advantage of C-Net and INet CNN architectures. INet is a CNN architecture that can increase receptive areas by incrementally increasing the kernel sizes of convolutional layers from (3 × 3) to (7 × 7) and then (15 × 15) without downsampling [12]. However, the convolutional layer of the INet architecture (3 × 3) is partially used in this study. Fig. 1 shows the proposed brain stroke diagnosis CNN scheme by combination of the C-Net and INet architectures.

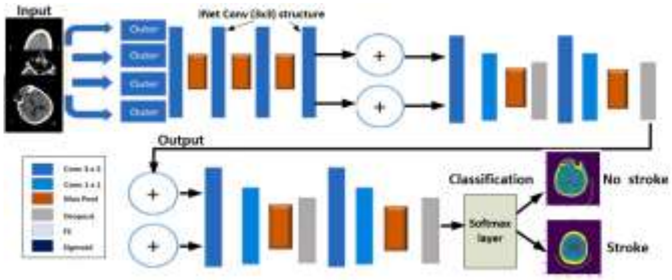


Fig. 1. The proposed hybrid CNN scheme for brain stroke diagnosis

On the C-Net side of the proposed model, various layers and parameters are designed to achieve various important goals, such as brain CT images classification and segmentation. C-Net is a CNN architecture consisting of Outer, Middle, and Inner networks. Using the same deep model in Outer networks provides a more stable and reliable way of feature extraction. Also, using the same (3 × 3) filter size in convolution layers of INet provides a better way of feature extraction compared to different operations with different filter sizes. INet (3 × 3) Conv layer structure of the proposed method is represented in Fig. (2), and explained later. Because deep networks operate in parallel, features are extracted by different networks at different times rather than directly connecting to fully connected (FC) layers, and as a result, the shortcomings of one network are compensated for the other network. The proposed model has few parameters with an image size of 256 × 256 pixels. In the model of each of the Outer networks, first, brain images enter the input layer in all Outer networks simultaneously. After that, it is processed through several convolution layers that do the convolution task represented as (1).

$$(X * K)(i,j) = \sum_m \sum_n \sum_c X(m,n,c) K(m+i-1, j+n-1, c) \quad (1)$$

where X defines a three-dimensional brain image that is being convolved by three-dimensional kernel K, sliding over all spatial positions. The first block has a maximum pooling layer. The block here consists of several convolutional layers followed by a pooling layer. The number of filters in the first block is 64 with (3×3) filter size and the same filling, the max-pooling filter size is (2×2), and it is repeated in 2 steps for three additional blocks with the same structure and the same order. The number of filters is multiplied by 2 excluding the last block. The maximum pooling layer for the last block has been reduced to prevent further reduction of the final output [15]. Rectified Linear Unite (ReLU) activation function is utilized in the convolution layers as defined in (2).

$$g_{m,n,c} = \max(0, w_c^T x_{m,n}) \quad (2)$$

where (m,n) denotes the parameters for the feature map, c denotes the channel index, w denotes the filter, and $x_{m,n}$ indicates the input at location (m,n). The returning features of the output are then concatenated, as represented by $\oplus$ operator in Fig. 1. This f operation is applied two by two on all of the output of the Outer networks as defined in (3).

$$f(y,w) = \left(\left(y_{m,n,c_i} \right) \oplus \left(w_{m,n,c_j} \right) \right) = X_{y_{m,n,c_i} + c_j} \quad (3)$$

where y and w denote feature maps of different networks, $(c_i\ ,\ c_j\ )$ denotes the number of channels in each output. Input of Middle networks are features extracted from Outer networks. Middle networks consist of four overlapping convolution layers, each with a (3 × 3) filter size, the same padding, and 256 filters. A (1 × 1) convolution is placed on top of previous convolutions to reduce the complexity of the model and feature maps. The feature maps obtained from the Middle networks are concatenated as in Eq. (3). This concatenation acts as the input for the Inner network. In this way, it is ensured that each network has the maximum pooling layer when generating efficient feature descriptors. Dropping, which randomly closes some components of the layers, is a normalization method. It also has the feature of preventing the network from over-learning. This process has been applied to each block of the Middle networks. Finally, the Inner network takes as input the features came back by the Middle networks. The Inner network consists of two convolution layers with filter size (3 × 3) and step 1, a block with the same padding and 256 filters. Also, the proposed model has a (1 × 1) convolution layer with the same structures, a size (2 × 2), and a 2-step maximum pooling layer. ReLU is utilized as the activation function in the Inner network. The inner network's maximum pooling layer is converted into a vector and coupled to an FC layer. After that, it is connected to another FC layer with an equal number of units. Both FC layers have been subjected to a release treatment. Finally, the sigmoid activation function has been used as shown in (4) and (5), respectively [15].

$$z = w^T x + b \quad (4)$$

$$\hat{y} = Sig(z) = \frac{e^z}{e^z + 1} \quad (5)$$

where z denotes the dot product of filter w with a part of the image with same filter size, and b denotes the bias. The loss function is a cross-entropy function given as (6).

$$L(\hat{y}, y) = -\left(\sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right) \quad (6)$$

where y_i denotes the ith label y of N classes, and $\hat{y}_i$ denotes the ith element of output $\hat{y}$. A total time complexity of the convolutional layers is computed as in (7).

$$C\left(\sum_{i=1}^{l} n_{i-1}\ .\ f_i^2 . c_i\ .\ m_i^2 \right) \quad (7)$$

where n_{i-1} and c_i denote the number of the input channels and filter channels in the ith layer, respectively, f_i denotes the filter size, and m_i denotes the output feature map size. The time may be decreased by the factor $\frac{1}{4}$ by adding a pooling layer with the same stride with (2 × 2) in each layer.

In addition to the C-Net architecture, the proposed hybrid model also integrates the superior features of the INet architecture, such as residual shortcuts and convolution index features. In this model, it combines the output feature maps of all previous convolution layers to extract features against kernels of different sizes [12]. Besides, the large kernel deep network is suitable for biomedical image classification and segmentation [18]. As can be seen in Fig. 1, a softmax layer has been used in the brain image classification process. The softmax layer ensures that the deep network output is a class rather than a numerical prediction [5]. This softmax is used in accordance with the class concept because it is desired to detect a brain CT image as non-stroke or stroke. While representing a non-stroke or non-stroke class here, it is expected that there will be a probability value for each class in the output of the proposed model. That is, the input values given to softmax are a kind of non-normalized version of the prediction values. Therefore, the more classes there are, the more output is obtained. In this study, 2 classes are represented as output. The softmax probability calculation ($\overline{p_w}$) is calculated as (8).

$$\overline{p_w} = \frac{e^{u_w}}{\sum_k e^{u_k}} \tag{8}$$

For the normalization process, the data whose probability is to be calculated must be divided by the sum of all data. Thanks to this operation, the probability sums $\overline{p_w}$ that is the probability sum of all classes become 1. In order for the sum of u_w and u_k in the formula to comply with the probability axiom, both rules must be satisfied. There are two options for providing the first rule: taking it in absolute value or expressing it as exponential.

Fig. 2 shows the designed residual shortcuts and convolutional index of INet. The right-hand side of Fig. 2 shows the residual shortcuts. It denotes the underlying second-last Conv mapping a Conv-layer as $G_i(x_i)$ and $H_i[G_i(x_i)]$, respectively. The stacked Convs fit another mapping: $D_i(x_i) = G_i(x_i) - x_{i-1}$ and $F_i[G_i(x_i)] = H_i[G_i(x_i)] - x_i$, respectively. Once $D_i(x_i)$ covers (i.e., $G_i(x_i) = x_{i-1}$), INet method optimizes $F_i(x_{i-1}) = H_i(x_{i-1}) - x_i$ instead of skiping the last Conv layer. Also, the method performs identity mapping as the basic residual shortcut (i.e., $H_i[G_i(x_i)] = x_i$). The left-hand side of Fig. 2 shows the convolutional index $G_{i-1}(x_{i-1})$. This index lets INet to skip the Inner Convs between the second Conv-layer and the last third Conv-layer. It is considered the feature maps concatenation as giving equal importance to all preceding Conv-layers in INet [12], [22]. The convolutional index is a larger weight on the output feature maps of the previous Conv-layer that includes the highest level semantics. Heavily, Conv-index enables INet removes the feature maps concatenation. In the proposed CNN, it is considered a Conv-layer to be defined as (9) and (10).

$$G_i(x_i) = D_i(x_i) + x_{i-1} + G_{i-1}(x_{i-1}) \tag{9}$$

$$H_i(x_i) = F_i[G_i(x_i)] + x_i \tag{10}$$

As shown in Fig. 1, the output of the proposed deep learning network architecture consists of two nodes, no-stroke and stroke. As mentioned before, a softmax layer has been used as the classification method, and 2 classes have been defined. So, it is determined whether there is a stroke or not from the brain CT images.

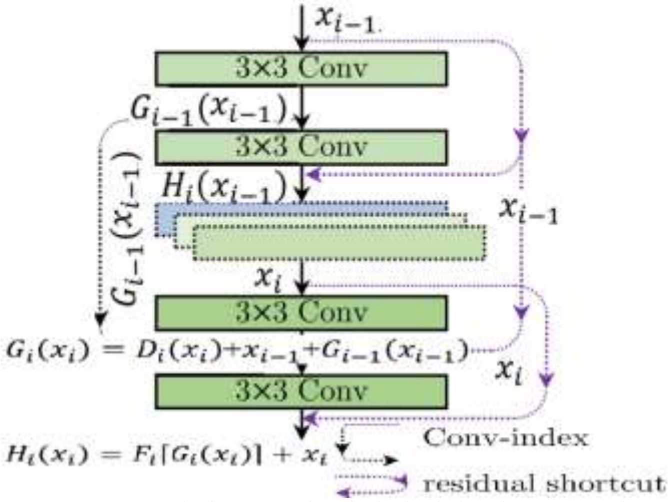


Fig. 2. INet structure in the proposed method. The designed (3 × 3) conv layers with residual shortcuts and convolutional index of INet

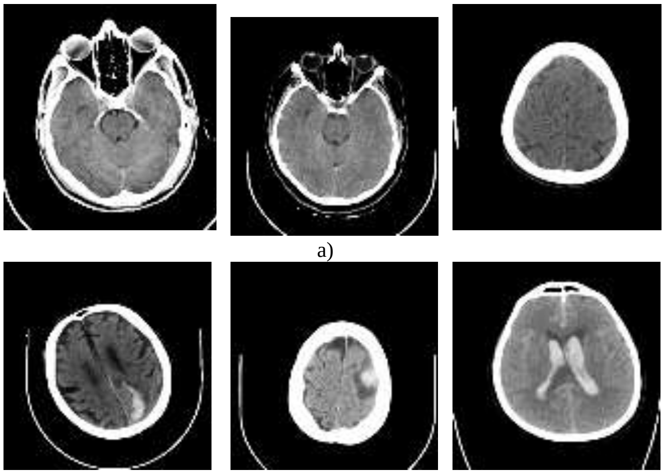
III. Experimental Analysis And Results

A. The Data Set Used in This Study

The Ministry of Health of the Republic of Turkey donated a data collection that included processed 256 × 256 pixel DICOM brain CT images. The data set contains 6650 CT brain scans, 4427 of which were stroke-free and 2223 of which were. To add more data to the brain imaging, data augmentation techniques have been used. It was given horizontal flipping and 20% rotation interval approaches, which stopped it from learning from unimportant characteristics and improved its performance as a whole [21]. To enhance the classification and segmentation performance even more, it has been decided to increase the amount of photos. The number of photographs with stroke has been doubled to 4446 because there are significantly fewer images with stroke than without. These strategies for data augmentation were used to add 80% of the data required for training and testing the classification model to the training set, and the remaining 20% was used for testing. Table 1 displays the number of brain CT scans utilized for training and testing in the dataset. In total, there are 7099 images in the training set of the classification model, of which 3542 are CT scans devoid of evidence of a stroke and 3557 are CT images containing such signs. Several CT scans of the brain are shown in Fig. 3. The photos in Fig. 3(a) are of some non-strokes. Images of various strokes caused by ischemia or hemorrhage are shown in Fig. 3(b).

TABLE I
THE NUMBERS OF BRAIN CT IMAGES FOR TRAINING AND TEST IN THE DATASET

Type of operation	**No stroke**	**Stroke**
For Training	3542	3557
For Test	885	889



a)

b)

Fig. 3. Some CT brain images of (a) no stroke (b) brain stroke in the dataset

B. Experimental Setup

The proposed method and other models have been implemented on Windows 10 operating system running Intel ®CoreTM i7-8700 processor and 16 GB RAM, Nvidia Geforce 4GB Graphics Card device using Python 3.8 programming for the experiments. Keras [23] and Tensorflow [24] libraries are utilized for training the proposed network. The deep learning framework is Pytorch 1.7.1 based on CUDA Tookit10.0. In the experiments, image preprocessing methods have been applied to achieve better image quality in classification and segmentation. Table 2 presents several parameters used in this study.

TABLE II
SEVERAL PARAMETERS USED IN THIS STUDY

Parameters	**Definition**
Convolution layer kernel size	(3 x 3) kernel size used
Output nodes	2 classes classification (no stroke or stroke)
Learning rate	0.001
Optimization method	Adam
Batch size	16
Number of epochs	100
Dropout	0.5

To evaluate the proposed model, precision (*Prc*), true positive rate (*Recall*), false positive rate (*FPR*), *F1-score*, and accuracy (*Acc*) as the evaluation metrics are defined, and calculated in (11), (12), (13), and (15), respectively [21].

$$Prc = \frac{TP}{TP+FP} \tag{11}$$

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

$$FPR = \frac{FP}{FP+TN} \tag{13}$$

$$F1-score = 2.\frac{PrcxRecall}{Prc+Recall} \tag{14}$$

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

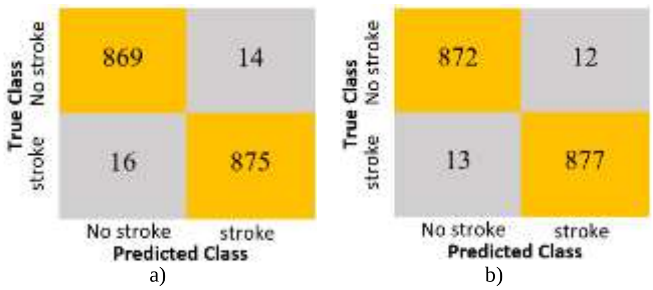
$$MCC = \frac{(TPxTN)-(FPxFN)}{\sqrt{(TP+FN)x(TP+FP)x(FP+TN)x(TN+FN)}} \tag{16}$$

where TP, TN, FP, and FN denote the true positive, true negative, false positive, and false negative, respectively. Also, Matthews Correlation Coefficient (MCC) is evaluated as given in Eq. (16). The MCC generates a high score when the network model has gained superior performance on all the groups of confusion matrix including TN, TP, FN, and FP, as more unbiased and reliable than the accuracy.

C. The Results of Brain Stroke Classification and Segmentation

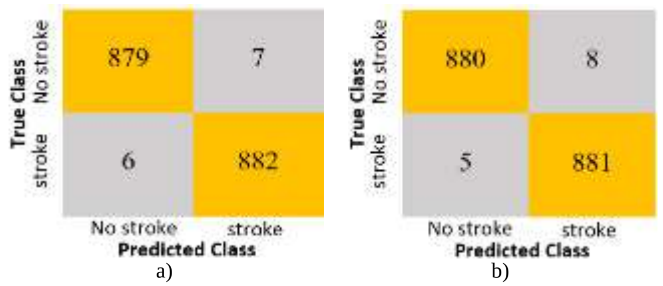
This section evaluates the brain CT images classification and segmentation results. The results of classification as no stroke or stroke in an image from brain images are presented, and then if there is a stroke in an image, the region of the image with the stroke is segmented.

Fig. 4 shows the confusion matrices for ResNet50v2 and UNet, Fig. 5 shows the confusion matrices for DeepLabV3 and ResUNet, Fig. 6 shows the confusion matrices for DenseINet and ResDenseUNet. The confusion matrices for C-Net and the proposed model are shown in Fig. 7 as experimental analysis findings for the categorization of stroke in brain CT images. The true class and the expected class are the two classes in which the confusion matrices are indicated. With this classification, it has been found how many of the brain images in the form of non-stroke or stroke have been estimated correctly. It is clearly seen that the least number of mistakes in stroke predictions are performed by using the proposed method. Table 3 presents the brain stroke classification performance results calculated from the confusion matrices. From the results of method performances, it is clearly deduced that the proposed method, C-Net, ResDenseUNet, DenseINet, ResUNet, DeepLabV3, UNet, and ResNet50v2 achieved at 99.43%, 99.32%, 99.1%, 99.21%, 99.1%, 98.87%, 98.65%, and 98.42% in precision, 99.66%, 99.43%, 99.43%, 99.32%, 99.1%, 98.76%, 98.53%, and 98.2% in recall, 99.54%, 99.37%, 99.26%, 99.26%, 99.1%, 98.81%, 98.58%, and 98.3% in F1-score, 99.54%, 99.37%, 99.26%, 99.26%, 99.09%, 98.81%, 98.59%, and 98.3% in accuracy, 99.1%, 98.76%, 99.53%, 98.53%, 98.2%, 97.63%, 97.18%, and 96.62% in MCC performances, respectively. Fig. 8 shows the accuracy results of the proposed model. It is understood that the accuracy rate of the proposed model is 99.54% after 100 epochs are completed. Fig. 9 shows the loss results of the proposed model. It is clearly understood that the loss rate of the proposed model is very low, almost zero. According to all these classification results, the best performances are obtained with the proposed model, and the proposed model outperformed other methods in terms of various benchmarks.



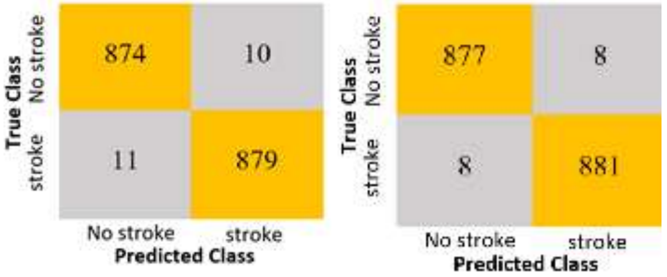
a) b)

Fig. 4. Confusion matrix of the brain stroke classification results using a) ResNet50v2, b) UNet



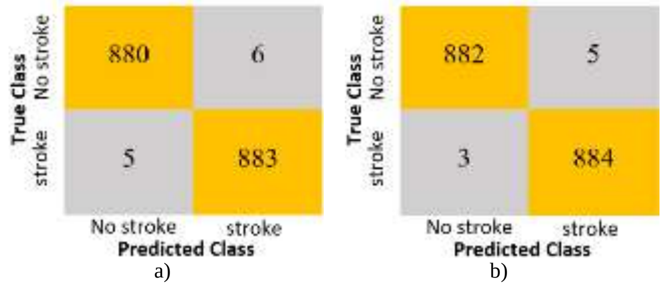
a) b)

Fig. 6. Confusion matrix of the brain stroke classification results using a) DenseINet, b) ResDenseUNet.



a) b)

Fig. 5. Confusion matrix of the brain stroke classification results using a) DeepLabV3, b) ResUNet.



a) b)

Fig. 7. Confusion matrix of the brain stroke classification results using a) C-Net b) Proposed Model

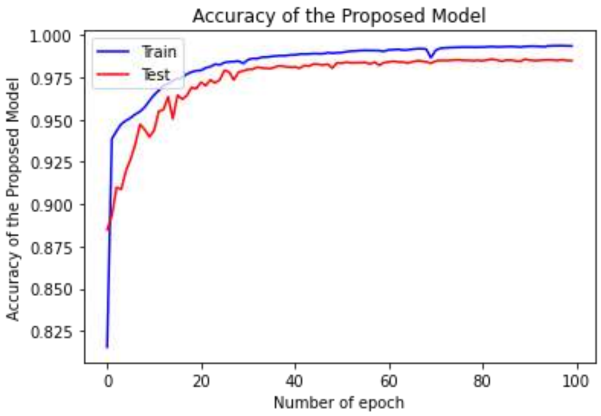


Fig. 8. Accuracy of the proposed model

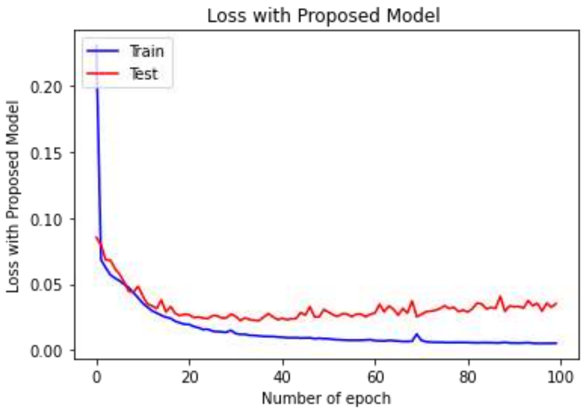


Fig. 9. Loss of the proposed model

TABLE III
PERFORMANCE RESULTS OF THE BRAIN STROKE CLASSIFICATION

Authors	**Method**	**Prc**	**Recall**	**FPR**	**F1-score**	**Acc**	**MCC**
Rahimzadeh and Attar (2020)	ResNet50v2	0.9842	0.9820	0.0158	0.9830	0.9830	0.9662
Ranneberger *et al.* (2015)	UNet	0.9865	0.9853	0.0135	0.9858	0.9859	0.9718
Chen et al. (2017)	DeepLabV3	0.9887	0.9876	0.0113	0.9881	0.9881	0.9763
Zhan et *al.* (2018)	ResUNet	0.9910	0.9910	0.0090	0.9910	0.9909	0.9820
Weng and Zhu (2021)	DenseINet	0.9921	0.9932	0.0079	0.9926	0.9926	0.9853
Khened *et al.* (2019)	ResDenseUNet	0.9910	0.9943	0.0090	0.9926	0.9926	0.9853
Barzekar and Yu (2022)	C-Net	0.9932	0.9943	0.0067	0.9937	0.9937	0.9876
	Proposed Model	**0.9943**	**0.9966**	**0.0056**	**0.9954**	**0.9954**	**0.9910**

The segmentation is made as follows. Images are divided a visual input into segments to make image analysis easier. Segments consist of one or more sets of pixels. While brain stroke segmentation breaks down pixels into larger components, there is also no need to view each pixel as a unit. It is the process of dividing an image into endurable segments or tiles. The stroke segmentation process begins with the identification of small regions on an image that should not be split. These areas are called strokes, and the position of these seeds defines the tiles. Fig. 10 illustrates the segmentation estimation results of brain images. Fig. 10(a) shows the original brain CT image. If there is a brain stroke from this overlay image, the stroke region is detected and segmented. The presence of overlay images (ground truth) in Fig. 10(b) indicates that radiologists have confirmed stroke images. The stroke area obtained from the stroke brain images approved by the radiologists is scanned in red. In Fig.10(c), the proposed model result is obtained, and the brain stroke regions scanned in green are given. As can be seen from these results, the estimation of stroke from brain CT images and the detection of the borders of the stroke region are quite successful. In addition, results very similar to the stroke areas determined by radiologists are obtained that are available in the dataset. Fig. 11 shows some estimations and segmentations of brain stroke. Here, some brain images with size of 256 x 256 are masked to make them dynamic with contrast, and strokes from masked images have been more easily detected.

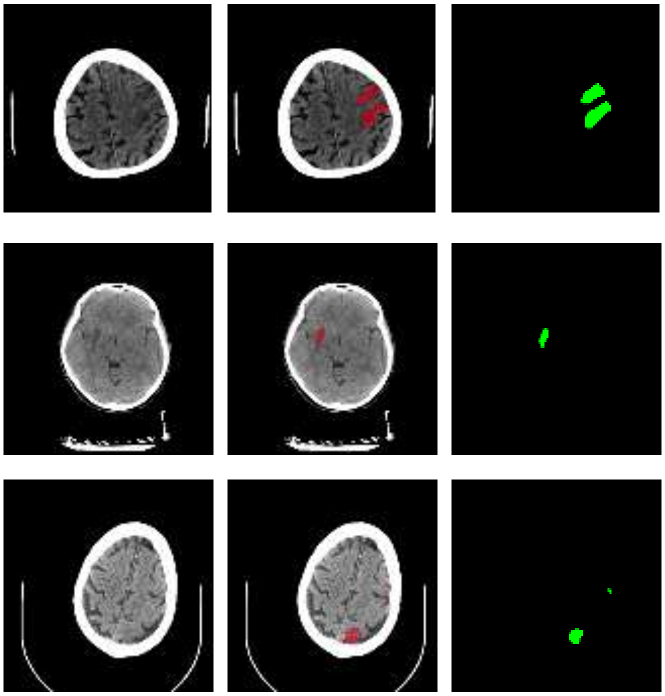

a) b) c)

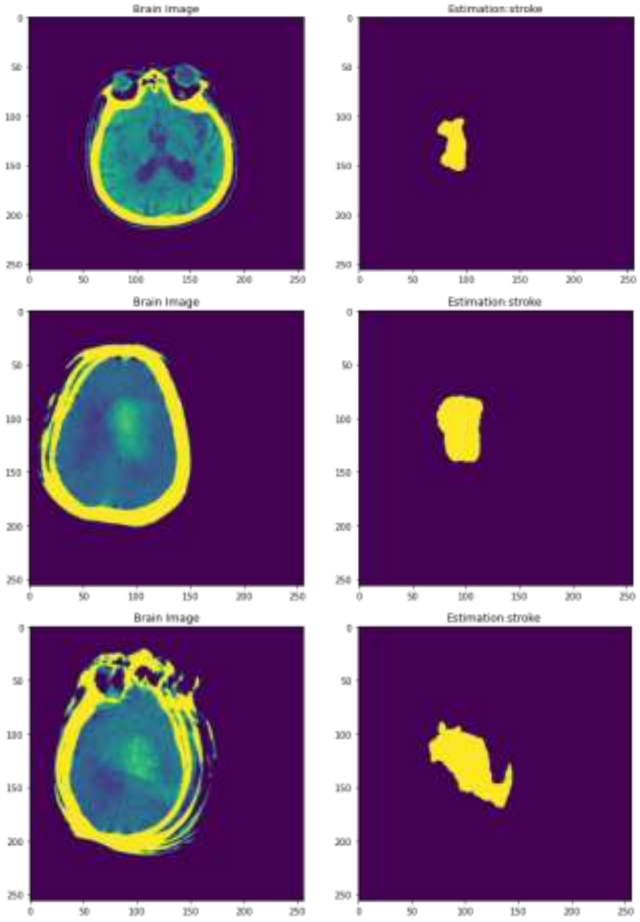
Fig. 10. Brain stroke estimations and segmentations a) original images b) stroke images confirmed by radiologists (overlay images) c) estimated and segmented images

Note that ten experts annotated ground truth maps for the evaluation of segmentation performance. The results of the segmentation are approved by 10 highly experienced radiologists. The Intersection over Union (IoU) and Dice Coefficient (DC) values are measured to prove the accuracy of the segmentation process.

Here, we want to evaluate how well spatial segmentation zones are predicted from photos of brain strokes. The IoU and DC computations are shown in Fig. 12. The IoU value is determined as a percentage of the number of pixels that are different from 0 at the intersection of the M_p and M_d pictures, as well as the intersection of the M_p and M_e images, as in (17).

$$IoU = \frac{S(Rectangle1 \cap Rectangle2)}{S(Rectangle1 \cup Rectangle2)} \qquad (17)$$

where M_d signifies the image acquired by the dilation operation using the 3x3 convolution matrix of the mask image, M_e means the image obtained by the erosion operation using the 3x3 convolution matrix of the mask image, and M_p defines the picture produced by the segmentation model. Each image's IoU value is determined separately, and when evaluating the models, the average of these values was taken into consideration. Rectangles 1 and 2 are presumptively represented by [x1, y1, x2, y2] and [x3, y3, x4, y4], respectively. The stroke segmentation zones are calculated using this convention.



a) b)

Fig. 11. Some estimations and segmentations of brain stroke a) masked images b) estimated and segmented of stroke

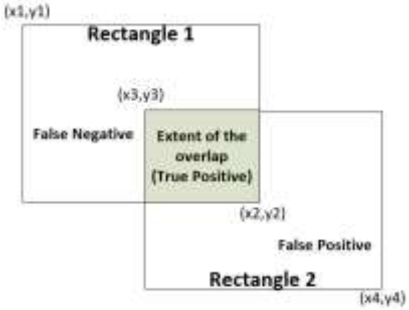


Fig. 12. Representation of the Intersection over Union (IoU) and Dice Coefficient (DC)

Accordingly, the Dice Coefficient (DC), which is also known as the Sørensen–Dice Coefficient, is defined as two times the area of the extent of the overlap, divided by the sum of the areas of Rectangle 1 and Rectangle 2 as given in (18):

$$DC = \frac{2x\ S(Rectangle1 \cap Rectangle2)}{S(Rectangle1) + S(Rectangle2)} \qquad (18)$$

The proposed model produces a higher IoU and DC accuracy rates and a lower loss rate in the experiments.

The average IoU and DC performance results are shown in Table 4. The proposed method, C-Net, ResDenseUNet, DenseINet, ResUNet, DeepLabV3, UNet, and ResNet50v2 achieved at 97.97%, 97.52%, 96.73%, 95.83%, 95.16%, 94.37%, 93.47%, and 93% in IoU, 98.97%, 98.74%, 98.34%, 97.87%, 97.52%, 97.1%, 96.62%, and 96.38% in DC, respectively.

TABLE IV
PERFORMANCE RESULTS OF THE BRAIN STROKE SEGMENTATION

Authors	**Method Name**	**Intersection-over-Union (IoU)**	**Dice Coefficient (DC)**
Rahimzadeh and Attar (2020)	ResNet50v2	0.93	0.9638
Ranneberger *et al.* (2015)	UNet	0.9347	0.9662
Chen et al. (2017)	DeepLabV3	0.9437	0.971
Zhan et *al.* (2018)	ResUNet	0.9516	0.9752
Weng and Zhu (2021)	DenseINet	0.9583	0.9787
Khened *et al.* (2019)	ResDenseUNet	0.9673	0.9834
Barzekar and Yu (2022)	C-Net	0.9752	0.9874
	Proposed Model	**0.9797**	**0.9897**

IV. Conclusion

In this paper, a brain stroke classification and segmentation method is proposed using C-Net and INet based CNN methods. Brain CT images are classified as no stroke or stroke using the proposed hybrid CNN model. Moreover, the proposed model finds the location and area of the brain stroke region with its boundaries by the segmentation method. To test the proposed model, several performance analyzes are performed with existing CNN methods such as ResNet50v2, UNet, DeepLabV3, ResUNet, DenseINet, ResDenseUNet, and C-Net using Python programming. From the performance results, it is concluded that the proposed model is better than other methods such as precision, recall, FPR, F1-score, accuracy, and MCC in classification and segmentation of the brain CT images. It is considered that the applying the proposed model for brain stroke diagnosis may be so useful for healthcare professionals in most medical applications. In future studies, it is planned to apply the proposed CNN model to the detection and classification of multiple diseases occurring in the abdomen or cardiovascular system.

V. ACKNOWLEDGMENT

The General Directorate of Health Information Systems of the Ministry of Health of the Republic of Turkey is thanked by the author for supplying the information for this study. According to Article 28 of the Law on Protection of Personal Data No. 6698 of the Republic of Turkey and Article 16 of the Regulation on Personal Health Data No. 30808, which was passed in accordance with the applicable law, this scientific investigation was conducted.

References

[1] S. Park, et al. “Annotated normal CT data of the abdomen for deep learning: Challenges and strategies for implementation”, Diagnostic and Interventional Imaging, Vol. 101, No. 1, 2020, pp.35-44.

[2] H. Huang, et al. “UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation”, Electrical Engineering and Systems Science, Image and Video Processing, 2020, https://doi.org/10.48550/arXiv.2004.08790.

[3] N. Dey, V. Rajinikanth, “Automated detection of ischemic stroke with brain MRI using machine learning and deep learning features”, Magnetic Resonance Imaging, Recording, Reconstruction and Assessment Primers in Biomedical Imaging Devices and Systems, 2022, pp.147-174.

[4] A. Gautam, B. Raman, “Towards effective classification of brain hemorrhagic and ischemic stroke using CNN”, Biomedical Signal Processing and Control, Vol. 63, Article 102178, 2021.

[5] B.R. Gaidhani, R. Rajamenakshi, S. Sonavane, “Brain Stroke Detection Using Convolutional Neural Network and Deep Learning Models”, 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, Sep 28-29, 2019, pp. 242-249.

[6] C.M. Lo, P.H. Hung, D.T. Lin, “Rapid Assessment of Acute Ischemic Stroke by Computed Tomography Using Deep Convolutional Neural Networks”, Journal of Digital Imaging, Vol. 34, 2021, pp. 637–646.

[7] N. Tomitaa, S. Jiangb, M.E. Maederc, S. Hassanpour, “Automatic post-stroke lesion segmentation on MR images using 3D residual convolutional neural network”, NeuroImage: Clinical, Vol. 27, 2020,102276.

[8] V. Badrinarayanan, A. Kendall, R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”, IEEE transactions on pattern analysis and machine intelligence, Vol. 39, No. 12, 2017, pp.2481-2495.

[9] O. Ronneberger, P. Fischer, T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 2015, pp.234-241.

[10] L. Liu, S. Chen, F. Zhang, F.X. Wu, Y. Pan, J. Wang, “Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI”, Neural Computing and Applications, Vol. 32, 2020, pp.6545–6558.

[11] Z. Zhang, Q. Liu, Y. Wang, “Road extraction by deep residual UNet,’’ IEEE Geosci. Remote Sens. Lett., Vol. 15, No. 5, pp. 749–753, May 2018.

[12] W. Weng, X. Zhu, “INet: Convolutional Networks for Biomedical Image Segmentation”, IEEE Access, Vol. 9, 2021, pp.16591-16603.

[13] M. Khened, V. A. Kollerathu, G. Krishnamurthi, “Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers”, Med. Image Anal., Vol. 51, pp. 21–45, Jan. 2019.

[14] G.S. Saragih, et al. “Ischemic Stroke Classification using Random Forests Based on Feature Extraction of Convolutional Neural Networks”, International Journal on Advanced Science Engineering Information Technology, Vol. 10, No. 5, 2020, pp.2177-2182.

[15] H Barzekar, Z. Yu, “C-Net: A reliable convolutional neural network for biomedical image classification”, Expert Systems With Applications, Vol. 187, 2022, 116003.

[16] I. Oksuz, “Brain MRI artefact detection and correction using convolutional neural networks”, Computer Methods and Programs in Biomedicine, Vol. 199, 2021, 105909.

[17] A. Deshpande et al. " Automatic segmentation, feature extraction and comparison of healthy and stroke cerebral vasculature", NeuroImage: Clinical, Vol. 30, 2021, 102573.

[18] Y. Zhou, W. Huang, P. Dong, Y. Xi, S. Wang, "D-UNet: A Dimension-Fusion U Shape Network for Chronic Stroke Lesion Segmentation", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 18, No. 3, 2021, pp.940-950.

[19] M. Rahimzadeh, A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2", Informatics in Medicine Unlocked, Vol. 19, 2020, 100360.

[20] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, "Rethinking atrous convolution for semantic image segmentation", arXiv:1706.05587. [Online]. 2017, Available: http://arxiv.org/abs/1706.05587.

[21] S. Yalçın, H. Vural, "Brain stroke classification and segmentation using encoder-decoder based deep convolutional neural networks", Computers in Biology and Medicine, Vol. 149, 2022, 105941.

[22] K.S. A. Kumara, A.Y. Prasad, J. Metan, "A hybrid deep CNN-Cov-19-Res-Net Transfer learning architype for an enhanced Brain tumor Detection and Classification scheme in medical image processing", Biomedical Signal Processing and Control, Vol. 76, 2022, 103631.

[23] F. Chollet, et al. Keras. https://github.com/fchollet/keras, 2015.

[24] M. Abadi, et al. "Tensorflow: A system for large-scale machine learning", In 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.

BIOGRAPHIES

Sercan YALÇIN received his BSc in Computer Engineering from Firat University, Turkey in 2013. He received his MSc and PhD in Computer Engineering, Firat University, Turkey in 2016 and 2021, respectively. He currently works at Adıyaman University. His research interests include wireless networks, data science, and artificial intelligence.

Alternative CPU and GPU Parallel Computing Approaches for Improving Sequential Analysis of Probability Associations in Short Texts

Dima Alnahas and Ahmet Arif Aydin

***Abstract*—In linguistics, probabilistic relation between co-occurrent words can provide useful interpretation of knowledge conveyed in a text. Connectivity patterns of vectorized representation of lexemes can be identified by using bigram models of word sequences. Similarity assessment of these patterns is performed by applying cosine similarity and mean squared error measures on word vectors of probabilistic relation matrix of text. Moreover, parallel computing is another important aspect for various domains that enables fast data processing and analytics. In this paper, we aim to demonstrate the benefit of parallel computing for computational challenges of extracting probabilistic relations between lexemes. In this study, we have explored performance limitations of sequential semantic similarity analysis and then implemented CPU and GPU parallel versions to show benefits of multicore CPU-GPU utilization for computationally demanding applications. Our results indicate that the alternative parallel computing implementations can be used to significantly enhance performance and applicability of probabilistic relation graph models in linguistic analyses.**

***Index Terms*—Text similarity, probability relations, parallel computing, CUDA, multicore processing, GPU.**

I. Introduction

PARALLEL COMPUTING is a very important concept in various domains for a variety of tasks since it aims efficient use of underlying hardware, decreasing processing time and saving existing resources. Many research efforts have been directed towards applying multiple computer resources to compute parallel versions of sequential tasks [1], [2], [3]. Moreover, various data processing tools such as Apache Spark , Apache Pig and Apache Hadoop [4] have been developed to provide parallel computing by utilizing Google's MapReduce paradigm. Furthermore, parallel processing is highly demanded in solving computing problems that require real-time solutions since the value of gleaned information is inversely proportional [5] with processing time for real-time data analytics [6]. Thus, spending less time is crucial to perform on-demand actions for fulfilling near real-time requests [7].

Due to increasing demand for computing power in Artificial Intelligence (AI), parallel computing has been applied in vast research efforts in AI and related research areas such as Natural Language Processing (NLP), Robotics, Machine Learning, Data Mining, etc. [8]. Numerous NLP techniques facilitate information retrieval and textual patterns analyses in short texts [9] with the purpose of enhancing the potential of semantic relations extraction and its applications in Linguistics. Moreover, in natural language processing, parallel processing techniques were proven to be effective for enhancing performance of applications such as lexical analysis and shallow parsing [10]. In lexical analysis, a probabilistic graph model can be a useful tool to analyze relations among word sequences in a given text since it allows representation of these relations with low complexity. Thus, this model can be attainable by calculating words co-occurrence probabilities, which can convey semantic features and grammatical structure of text while reducing repeated lexical relations. Further exploration of language characteristics obtained from probabilistic associations of lexemes can provide more insights on relational similarity among words in short texts. In addition, vector representation of probabilistic associations among words allows for utilization of Cosine Similarity (CS) and Mean Squared Error (MSE) measures to perform relational similarity analysis [11].

One of the main purposes of this study is to apply parallel computing to perform fast and scalable relational similarity analysis on short texts of various lengths. In this paper, first, a sequential version of relational similarity analysis has been implemented. The computational complexity of relational similarity analysis and probabilistic graph model increases in proportion to the number of lexemes in the given text hence nodes in graph model and consequently imposes additional cost on time requirement of traditional serial computing. Next, one CPU and one GPU parallel versions have been implemented to get benefit of parallel computing and to decrease data processing time. These parallel versions provide a significant decrease in run time required to perform relational similarity analysis regarding the sequential version. Last, performance evaluations are presented by comparing the sequential version with the proposed CPU and GPU parallel computing approaches on the same text data.

This paper is organized as follows. In section II, a related work is presented and in section III, the methodology of the probabilistic relation graph model is explained. Then,

Dima Alnahas is with the Department of R&D, Infina Software Inc., Istanbul, TURKEY e-mail:dalnahas@infina.com.tr

Ahmet Arif Aydin is with the Department of Computer Engineering, Engineering Faculty, Inonu University, Malatya, 44000 TURKEY e-mail: arif.aydin@inonu.edu.tr

Manuscript received Feb 07, 2022; accepted Sep 30, 2022.
DOI: 10.17694/bajece.1069152

in section IV, our implementations of the sequential and two parallel version approaches are presented. In section V, computational comparisons and results of our implementations and evaluations are provided and in section VI, a conclusion is provided to present the contributions of our work.

II. Related Work

In many studies, co-occurrence probabilities of lexemes and its vector representation have been utilized in natural language processing. Also, Statistical Language Modeling is considered a successful approach for various tasks of NLP such as, machine translation, text classification, spelling correction, etc. However, similar approaches involve intensive matrix computations and analyses, thus, requiring immense computation time, power usage, and resources.

The first use of word co-occurrence probabilities in language modeling dates back to 1999. I. Dagan et al. [12] utilized a probabilistic word association model in tasks of language modeling and pseudo-word disambiguation.

In a recent work, A. Schakel and B.J. Wilson [13] introduced the use of word co-occurrence and vector representation as a significant factor of word in corpus. This study further explores the language features that can be conveyed by Word2vec [14]. In a later study, D. Alnahas and B.B. Alagoz [11] suggested a deep relational similarity analysis which explores path probabilities between words by utilizing power of probabilistic relation matrix. More recently, Y. Yin et al. [15] introduced a method to improve accuracy of text recommendation by 8.63%. The method in [15] utilizes improved cosine similarity measure to compare correlation coefficients vectors of related texts.

Moreover, in an attempt to minimize the computational cost, Mikolov et al. [16] presented the Skip-gram Model which utilizes probability to predict surrounding words in a short text. This study suggests training the Skip-gram model with distributed representations of words as a solution to achieve learned representation of phrases with minimal computational complexity.

In [17], authors accelerated text clustering speed while performing text similarity measurement by utilizing Spark architecture in parallel computing. In further effort to minimize computational cost, many researchers have explored the possible utilization of CPU and GPU cores. In a performance analysis study, S. Gupta and M.R. Babu [10] demonstrated that a 16-core GPU performs expectedly better than single-core and multi-core CPUs in the simple task of string matching. Furthermore, in a more recent study, E. Strubell et al. [18] described the financial and environmental impact of training state-of-the-art NLP models using large computational resources. This study compares the carbon emissions from training common NLP models to familiar consumptions such as, Air travel. As a result, E. Strubell et al emphasizes the need for NLP models that can be trained and developed on more affordable computational resources such as commodity laptop or server, while providing state-of-the-art analysis. As a result, these studies inspired our research to provide faster alternatives to existing NLP models and training methods by efficiently applying available hardware resources in similarity-based NLP analyses.

III. Methodology

In this section, first, a probabilistic relation graph model approach is presented for relational similarity analysis of short texts. Then, two similarity measures are applied to evaluate similarity level of probabilistic relations of word pairs. Last, an illustrative example is provided to demonstrate probabilistic associations analysis in short text.

A. Probabilistic Relation Graph Model of Short Text for Semantic Similarity Analysis

In linguistics, a word sequence of finite length can be interpreted to a form of knowledge or information. Also, word sequences can be depicted as messages and a series of messages represents a text. A vocabulary set of a message collection consists of lexical instances of message elements. Adjacent words in a message are considered to have bigram relation. The bigram relation frequency matrix of co-occurrent word pairs conveys information of co-occurrence frequency of words instances in vocabulary set. Let us form a vocabulary set of message series $M_1, M_2, .., M_h$ as,

$$W_c = w_i : w_i \perp M_1 \vee w_i \perp M_2 \vee ... \vee w_i \perp M_h \,, \tag{1}$$

where the occurrence operator $\perp$ infers that w_i item is an element of M_j message in the term $w_i \perp M_j$. The bigram relation frequency matrix of a message M is constructed by

$$R_f = \begin{cases} f_{i,j} = f_{i,j} + 1 \,, & ''w_i w_j'' \perp M \wedge w_i, w_j \in W_c \\ f_{i,j} = f_{i,j} & otherwise \end{cases} \tag{2}$$

Probabilistic associations between co-occurrent lexeme pairs in finite-length word sequences can be expressed by a probabilistic bigram relation graph model. The probabilistic relation matrix of bigram model is identified as normalized values of R_f elements in a range of $[0, 1]$,

$$R_p \cong \frac{1}{\sum R_f} R_f \,, \tag{3}$$

where $\sum R_f$ is summation of R_f elements and is calculated by $\sum_{i=1,j=1}^{k,k} f_{i,j}$. Each element of probabilistic relation matrix conveys the probability of relation between i^{th} and j^{th} elements of vocabulary set. Accuracy of estimating the co-occurrence probability of two lexeme items increases in proportion to length of text.

Fig. 1 illustrates a message example of word sequence $M = ''w_1 w_2 w_3 w_4 w_2 w_4''$. In this figure, probabilistic transitions between co-occurrent lexeme items in word sequence M are demonstrated by probabilistic relation matrix R_p and its corresponding weighted graph representation. Zero value of probabilistic relation matrix element $p_{i,j}$ depicts the absence of co-occurrence between w_i and w_j items of text.

$$R_p = \begin{bmatrix} 0 & p_{1,2} & 0 & 0 \\ 0 & 0 & p_{2,3} & p_{2,4} \\ 0 & 0 & 0 & p_{3,4} \\ 0 & p_{4,2} & 0 & 0 \end{bmatrix}$$

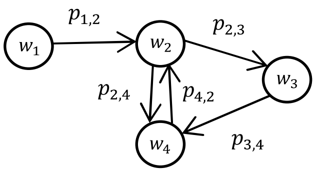


Fig. 1. A probability relation matrix of M message and corresponding weighted graph representation.

As demonstrated in Fig. 1, on one hand, the output-word vector of node w_2 is described by the 2^{nd} row elements of probabilistic matrix and is expressed as,
$u_2 = [0 \quad 0 \quad p_{2,3} \quad p_{2,4}]$.
On the other hand, the input word-vector of node w_2 is described by the 2^{nd} column elements of probabilistic matrix and is expressed as,
$v_2 = [p_{1,2} \quad 0 \quad 0 \quad p_{4,2}]^T$.
As a result, input and output word-vectors are utilized to evaluate similarity of relational transition paths between nodes. Therefore, similar probabilistic connection patterns can be an indication of relational similarity of a word pair.

B. Relational Similarity Measures Application on Vectorized Representation of Lexemes

In this study, Cosine Similarity (CS) measure is applied to assess relational similarity of lexeme pairs. CS matrix of output word-vectors can be obtained for all lexeme pairs of vocabulary set and is expressed as,

$$C_u = R_p \otimes R_p^T \, , \tag{4}$$

where the CS operator $\otimes$ performs CS calculation between vectors of R_p and R_p^T. CS matrix of input word-vectors can also be obtained for all lexeme pairs of vocabulary set and is expressed as,

$$C_v = R_p^T \otimes R_p \, . \tag{5}$$

Relational similarity based on CS is calculated with formulas 4 and 5 and can be expressed and normalized to the range of $[0, 1]$ as follows

$$C = \frac{1}{2}(C_u + C_v) \, . \tag{6}$$

Another measure to evaluate relational similarity of lexeme pairs is Mean Squared Error (MSE). Output MSE matrix is calculated for all lexemes in a vocabulary and can be expressed as,

$$E_u = R_p \ominus R_p^T \, , \tag{7}$$

where operator $\ominus$ performs MSE calculation between vectors of R_p and R_p^T. Input MSE matrix is calculated for all lexemes in vocabulary and can be expressed as,

$$E_v = R_p^T \ominus R_p \, , \tag{8}$$

MSE matrix E is then calculated by using formulas 7 and 8 for lexeme pairs and is expressed as

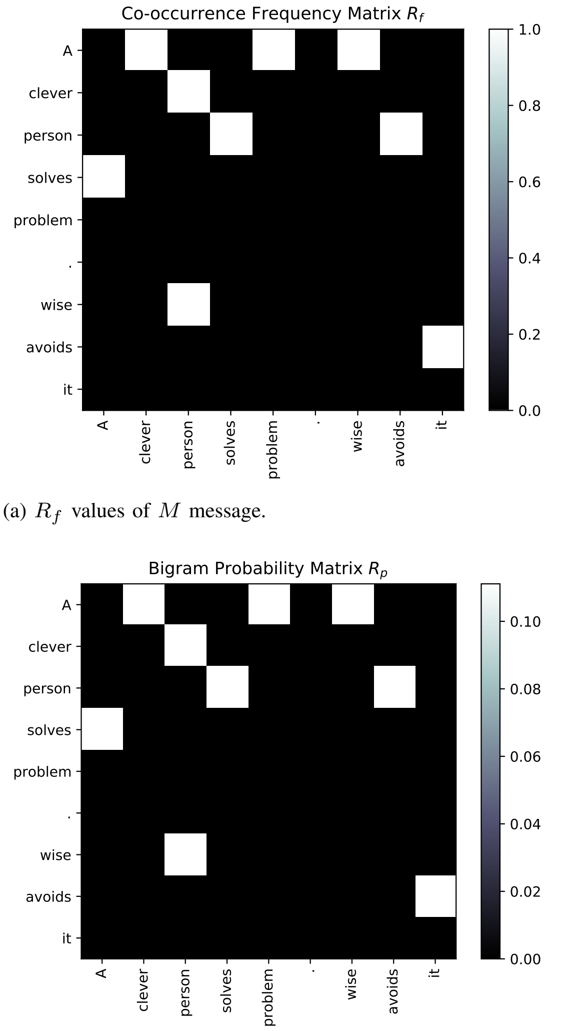
$$E = E_u + E_v \, . \tag{9}$$

C. An Explanatory Example of Probabilistic Relations Analysis in Short Text

Let us consider the following text which is a quote by Einstein:
M = “A clever person solves a problem. A wise person avoids it.”
This message provides the following vocabulary set:
W = A, clever, person, solves, problem, ., wise, avoids, it
Fig. 2a presents R_f matrix values of message M.
Fig. 2b shows values of R_p matrix as calculated using formula 3.
The full stop is assigned an index in the vocabulary set and it indicates the end of a sentence. As Fig. 2a demonstrates, the full stop is not included in the probabilistic calculations of associated lexemes. Fig. 3 presents bigram graph model of M. In this figure, the stream of transitions between lexemes is interrupted by full stop at the end of each sentence.
CS matrix of message M is illustrated in Fig. 4a that shows CS value of 1 for lexeme pair (wise-clever). This similarity measure indicates similar relations with similar adjacent words of the lexeme pair. In M message example, these similar adjacent words can be identified as “A” and “person” and similar relations can be detected in bigram relation graph of M.
Similarly, MSE matrix presents 0 error value for lexeme pair (wise-clever) which indicates validity of similarity analysis. The MSE matrix as calculated with formula 9 is demonstrated in Fig. 4b.
Diagonal values of CS and MSE matrices express similarity of each lexeme with itself which explains 1 values of diagonal in CS matrix and 0 values of diagonal in MSE matrix.

IV. Implementation

In this section, sequential, CPU parallel and GPU parallel versions are explained.



(a) R_f values of M message.

(b) R_p values of M message.

Fig. 2. R_f and R_p values of M message.

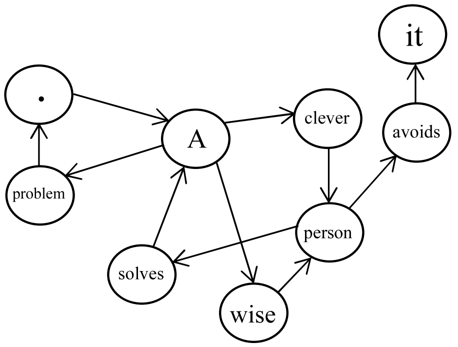
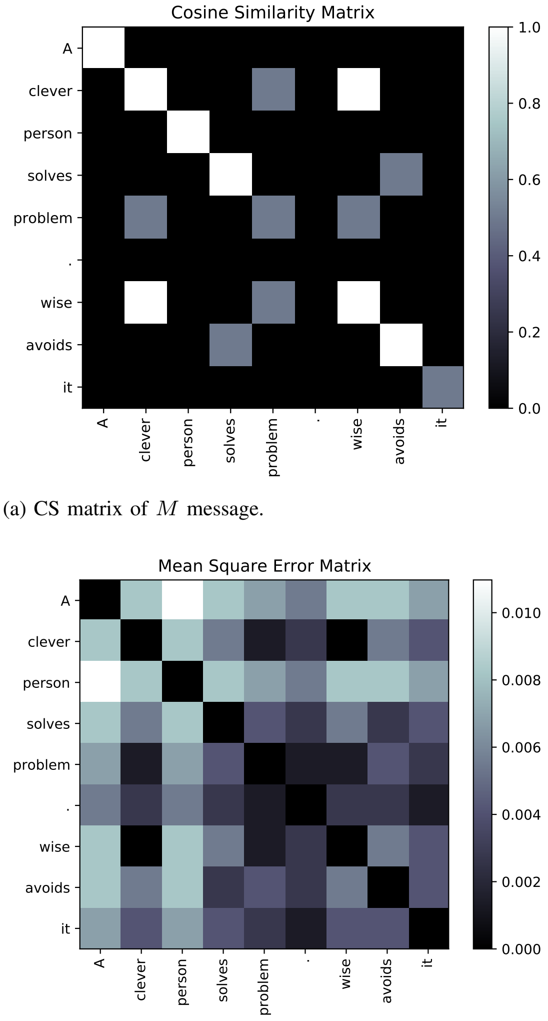


Fig. 3. Bigram relation graph of M message.



(a) CS matrix of M message.

(b) MSE matrix of M message.

Fig. 4. CS and MSE matrices of M message.

A. Sequential Probabilistic Similarity Analysis

Pseudocode of the sequential version of probabilistic similarity analysis algorithm is provided in Algorithm IV-A.

```
Pre-process text file
Create dictionary $W$ of text
for each element $i$ of $W$ do
    Obtain $l$ list of adjacent elements to $i$
    for each element $j$ of $l$ do
        $R_f(i,j) \leftarrow R_f(i,j) + 1$
Total frequency $T \leftarrow$ sum of $R_f$ elements
for each element $p$ of $R_p$ and corresponding element $f$ of $R_f$ do
    $p \leftarrow \frac{f}{T}$
Calculate Euclidean norm vector $N_u$ of $R_p$
Norms multiplication matrix $N_{u_{matrix}} \leftarrow N_u \times N_u^T$
for each $n_u$ element of $N_{u_{matrix}}$ do
    if $n_u = 0$ then
        $n_u \leftarrow 1$
for each element $u$ of $C_u$ and corresponding elements $n_u$ of $N_{u_{matrix}}$ , $p$ of $R_p$ and $p_T$ of $R_p^T$ do
```

```
  $u \leftarrow \frac{(p \times p_T)}{n_u}$
Calculate Euclidean norm vector $N_v$ of $R_p^T$
Norms multiplication matrix $N_{v_{matrix}} \leftarrow N_v \times N_v^T$
for each $n_v$ element of $N_{v_{matrix}}$ do
  if $n_v = 0$ then
    $n_v \leftarrow 1$
for each element $v$ of $C_v$ and corresponding elements $n_v$ of $N_{v_{matrix}}$ , $p$ of $R_p$ and $p_T$ of $R_p^T$ do
  $v \leftarrow \frac{(p \times p_T)}{n_v}$
for each element $c$ of $C$ and corresponding elements $u$ of $C_u$ and $v$ of $C_v$ do
  $c \leftarrow \frac{(u+v)}{2}$
Calculate $R_p$ size $S$
for each $e_u(x_u, y_u)$ element of $E_u$ and corresponding $x_u$ and $y_u$ vectors of $R_p$ do
  $e_u(x_u, y_u) \leftarrow$ sum of $(x_u - y_u)^2$ elements $/S$
for each $e_v(x_v, y_v)$ element of $E_v$ and corresponding $x_v$ and $y_v$ vectors of $R_p$ do
  $e_v(x_v, y_v) \leftarrow$ sum of $(x_v - y_v)^2$ elements $/S$
for each element $e$ of $E$ and corresponding elements $e_u$ of $E_u$ and $e_v$ of $E_v$ do
  $e \leftarrow e_u + e_v$
```

Fig. 5 illustrates serially computed matrices in the sequential version of the algorithm as explained in section III-C.

To obtain co-occurrence matrix R_f for probabilistic similarity analysis, words associations of M-length text are explored sequentially to construct a dictionary of lexeme items in text. For a dictionary of N elements, indexes of dictionary lexemes are used as columns and rows index of R_f. Therefore, R_f matrix conveys frequency information of $N \times N$ possible lexeme pair co-occurrences. Inspecting relational connectivity of word pairs with window size of 2 in M-length text requires $M - 1$ iterations in order to construct R_f matrix. To reduce computing time of frequently adjacent lexeme pairs in textual patterns such as grammatical structures, this study implies to alternatively iterate dictionary elements and obtain adjacent words list with corresponding occurrence frequency for each item in the dictionary. The inferred computing sequence can significantly reduce execution time particularly for long texts that produce relatively small vocabulary sets.

Probabilistic relation matrix R_p is populated by using R_f as expressed in formula 3. R_p describes association probability between each two items in dictionary, thus, it is obtainable by calculating co-occurrence probability of $N \times N$ word pairs.

CS matrices of input and output word-vectors of R_p and R_p^T are computed discretely to obtain CS matrix as expressed in formula 6. Each element of both CS matrices is a depiction of cosine the angle between input and output vectors of word pairs. Similarly, MSE matrices of input and output word-vectors of R_p and R_p^T requires sequential computing of Euclidean distance between input and output word-vectors of $N \times N$ word pairs.

B. Parallel Computing Approaches for Probabilistic Similarity Analysis

In this section, one CPU and one GPU parallel approaches are presented. These approaches have been developed to efficiently explore information extraction properties of connectivity analysis for long texts in parallel by making use of parallel computing tools. CPU and GPU parallel approaches are respectively explained next.

1) Multicore CPU Version: In this parallel CPU version, the Multiprocessing module of Python programming language is utilized to parallelize the process of extracting connectivity features of word sequences. The Multiprocessing module contains the Pool class which automatically initiates processes as many as core number of CPU when process number is not deliberately specified. Using the Map function, this class can divide elements of the argument matrix between the spawned processes which simultaneously execute the specified task by making use of underlying CPU cores. The computing steps of the CPU parallel version is provided in Fig. 6 that also presents how matrices are computed in this parallel version of probabilistic similarity analysis algorithm.

In this parallel CPU version, two main portions of the sequential algorithm have been parallelized. First, obtaining CS matrix for input and output word-vectors of R_p and R_p^T implemented by making use of Multiprocessing Pool class that drastically reduces time complexity of this task. Alternative to iterating matrix elements sequentially, elements are divided into chunks between initialized processes. Then, processes concurrently execute serial instructions of computing CS of lexeme pairs to construct CS matrix. The second main parallelized portion of the sequential algorithm is computing MSE matrix for input and output word-vectors of R_p and R_p^T. In this part word-vector pair chunks are distributed among processes which discretely calculate the MSE index of each pair.

2) A GPU Version: In this section, a GPU version of probabilistic similarity analysis is presented. Modern graphic processors have introduced drastic solutions for immensely large computation problems. Also, many research efforts continue to investigate the prospect of utilizing GPUs to reduce time complexity of analysis tasks in fields of deep learning and information processing [19]. In particular, CUDA (Compute Unified Device Architecture) programming model has been utilized to establish the required integration of NVIDIA's GPU with multicore CPU for parallel processing purposes. This parallel computing platform applies Single Instruction Multiple Thread (SIMT) execution model to manage and schedule warps independently, hence concurrently [20]. Moreover, instruction of threads can be specified by a C function that denotes a kernel that is executed concurrently by all available threads in the instruction sequence. Threads are organized in grids which consist of thread blocks. Each block is assigned a shared memory space that can be accessible by 512 threads within the block [21] and up to 1024 threads with recent CUDA toolkit.

In this paper, we aimed to explore CUDA's potential on concurrent extraction of textual features in near real-time manner. With the aid of CUDA, time complexity of obtaining similarity measures and probabilistic connectivity matrices of text is significantly reduced by efficiently expressing word

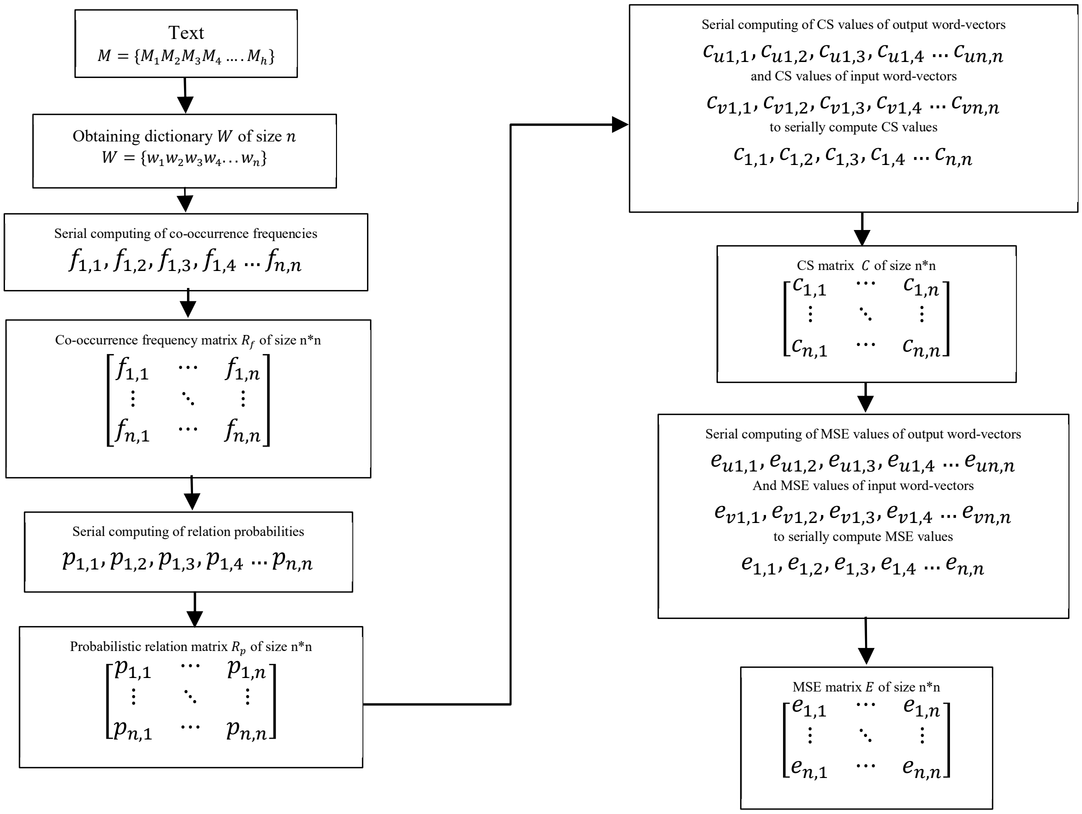

Text
$M = \{M_1 M_2 M_3 M_4 \ldots . M_h\}$
Obtaining dictionary W of size n
$W = \{w_1 w_2 w_3 w_4 \ldots w_n\}$
Serial computing of co-occurrence frequencies
$f_{1,1}, f_{1,2}, f_{1,3}, f_{1,4} \cdots f_{n,n}$
Co-occurrence frequency matrix R_f of size n*n
$\begin{bmatrix} f_{1,1} & \cdots & f_{1,n} \\ \vdots & \ddots & \vdots \\ f_{n,1} & \cdots & f_{n,n} \end{bmatrix}$
Serial computing of relation probabilities
$p_{1,1}, p_{1,2}, p_{1,3}, p_{1,4} \cdots p_{n,n}$
Probabilistic relation matrix R_p of size n*n
$\begin{bmatrix} p_{1,1} & \cdots & p_{1,n} \\ \vdots & \ddots & \vdots \\ p_{n,1} & \cdots & p_{n,n} \end{bmatrix}$
Serial computing of CS values of output word-vectors
$c_{u1,1}, c_{u1,2}, c_{u1,3}, c_{u1,4} \cdots c_{un,n}$
and CS values of input word-vectors
$c_{v1,1}, c_{v1,2}, c_{v1,3}, c_{v1,4} \cdots c_{vn,n}$
to serially compute CS values
$c_{1,1}, c_{1,2}, c_{1,3}, c_{1,4} \cdots c_{n,n}$
CS matrix C of size n*n
$\begin{bmatrix} c_{1,1} & \cdots & c_{1,n} \\ \vdots & \ddots & \vdots \\ c_{n,1} & \cdots & c_{n,n} \end{bmatrix}$
Serial computing of MSE values of output word-vectors
$e_{u1,1}, e_{u1,2}, e_{u1,3}, e_{u1,4} \cdots e_{un,n}$
And MSE values of input word-vectors
$e_{v1,1}, e_{v1,2}, e_{v1,3}, e_{v1,4} \cdots e_{vn,n}$
to serially compute MSE values
$e_{1,1}, e_{1,2}, e_{1,3}, e_{1,4} \cdots e_{n,n}$
MSE matrix E of size n*n
$\begin{bmatrix} e_{1,1} & \cdots & e_{1,n} \\ \vdots & \ddots & \vdots \\ e_{n,1} & \cdots & e_{n,n} \end{bmatrix}$

Fig. 5. Computing steps of the serial version of similarity analysis.

sequences in blocks. At least eight blocks per multiprocessor can be executed simultaneously in association with hardware limitations and memory resources.

Fig. 7 presents possible blocks division of processed matrices in GPU version of similarity analysis. Each element of these blocks corresponds to a thread which computes one element of the processed matrix. Block number is usually assigned in accordance to data size.

Pseudo code of the GPU version of probabilistic similarity analysis algorithm is presented in Algorithm IV-B2.

```
Pre-process text file
Create dictionary $W$ of text
for each element $i$ of $W$ do
    Obtain $l$ list of adjacent elements to $i$
    for each element $j$ of $l$ do
        $R_f(i,j) \leftarrow R_f(i,j)+1$
Total frequency $T \leftarrow$ sum of $R_f$ elements
Initialize block dimensions $B_{dim}(x_{thread}, y_{thread}, z_{thread})$, total thread number $t \leftarrow x_{thread}+y_{thread}+z_{thread}$
Initialize grid dimensions $G_{dim}(x_{block}, y_{block}, z_{block})$, total block number $t \leftarrow x_{block}+y_{block}+z_{block}$
Divide $R_p$ matrix to $b$ blocks
for each block in parallel do
    for each element $p$ of $R_p$ and corresponding element $f$ of $R_f$ do
        $p \leftarrow \frac{f}{T}$
Calculate Euclidean norm vector $N_u$ of $R_p$
Norms multiplication matrix $N_{u_{matrix}} \leftarrow N_u \times N_u^T$
Divide $N_{u_{matrix}}$ and $C_u$ matrices to $b$ blocks
for each block in parallel do
    for each $n_u$ element of $N_{u_{matrix}}$ block do
        if $n_u = 0$ then
            $n_u \leftarrow 1$
    for each element $u$ of $C_u$ and corresponding elements $n_u$ of $N_{u_{matrix}}$ block , $p$ of $R_p$ block and $p_T$ of $R_p^T$ block do
        $u \leftarrow \frac{(p \times p_T)}{n_u}$
Calculate Euclidean norm vector $N_v$ of $R_p^T$
Norms multiplication matrix $N_{v_{matrix}} \leftarrow N_v \times N_v^T$
Divide $N_{v_{matrix}}$ and $C_v$ matrices to $b$ blocks
for each block in parallel do
    for each $n_v$ element of $N_{v_{matrix}}$ block do
        if $n_v = 0$ then
            $n_v \leftarrow 1$
    for each element $v$ of $C_v$ and corresponding elements $n_v$ of $N_{v_{matrix}}$ block , $p$ of $R_p$ block and $p_T$ of $R_p^T$ block do
        $v \leftarrow \frac{(p \times p_T)}{n_v}$
for each element $c$ of $C$ and corresponding elements $u$ of $C_u$ and $v$ of $C_v$ do
    $c \leftarrow \frac{(u+v)}{2}$
Calculate $R_p$ size $S$
Divide $E_u$ to $b$ blocks
for each block in parallel do
    for each $e_u(x_u, y_u)$ element of $E_u$ block and corresponding $x_u$ and $y_u$ vectors of $R_p$ block do
        $e_u(x_u, y_u) \leftarrow$ sum of $(x_u - y_u)^2$ elements$/S$
Divide $E_v$ to $b$ blocks
for each block in parallel do
    for each $e_v(x_v, y_v)$ element of $E_v$ block and corresponding $x_v$ and $y_v$ vectors of $R_p$ block do
        $e_v(x_v, y_v) \leftarrow$ sum of $(x_v - y_v)^2$ elements$/S$
for each element $e$ of $E$ and corresponding elements $e_u$ of $E_u$ and $e_v$ of $E_v$ do
    $e \leftarrow e_u + e_v$
```

V. Results and Evaluations

In this section, performance evaluations of sequential and parallel versions of probabilistic similarity analysis approaches are presented. In addition, Fig. 8 demonstrates performance

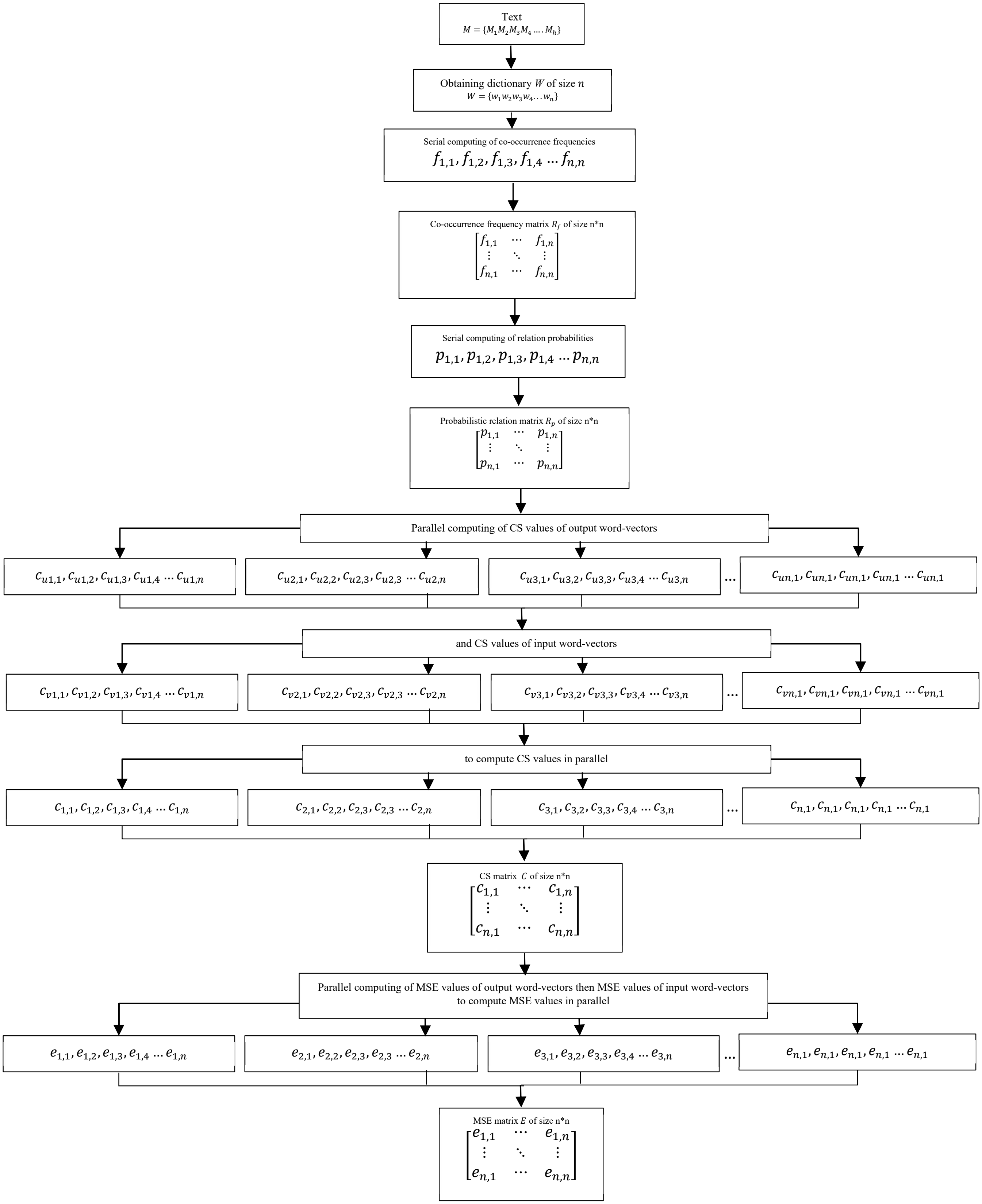

Text
$M = \{M_1 M_2 M_3 M_4 \ldots . M_h\}$
Obtaining dictionary W of size n
$W = \{w_1 w_2 w_3 w_4 \ldots w_n\}$
Serial computing of co-occurrence frequencies
$f_{1,1}, f_{1,2}, f_{1,3}, f_{1,4} \cdots f_{n,n}$
Co-occurrence frequency matrix R_f of size n*n
$\begin{bmatrix} f_{1,1} & \cdots & f_{1,n} \\ \vdots & \ddots & \vdots \\ f_{n,1} & \cdots & f_{n,n} \end{bmatrix}$
Serial computing of relation probabilities
$p_{1,1}, p_{1,2}, p_{1,3}, p_{1,4} \cdots p_{n,n}$
Probabilistic relation matrix R_p of size n*n
$\begin{bmatrix} p_{1,1} & \cdots & p_{1,n} \\ \vdots & \ddots & \vdots \\ p_{n,1} & \cdots & p_{n,n} \end{bmatrix}$
Parallel computing of CS values of output word-vectors
$c_{u1,1}, c_{u1,2}, c_{u1,3}, c_{u1,4} \cdots c_{u1,n}$
$c_{u2,1}, c_{u2,2}, c_{u2,3}, c_{u2,3} \cdots c_{u2,n}$
$c_{u3,1}, c_{u3,2}, c_{u3,3}, c_{u3,4} \cdots c_{u3,n}$
…
$c_{un,1}, c_{un,1}, c_{un,1}, c_{un,1} \cdots c_{un,1}$
and CS values of input word-vectors
$c_{v1,1}, c_{v1,2}, c_{v1,3}, c_{v1,4} \cdots c_{v1,n}$
$c_{v2,1}, c_{v2,2}, c_{v2,3}, c_{v2,3} \cdots c_{v2,n}$
$c_{v3,1}, c_{v3,2}, c_{v3,3}, c_{v3,4} \cdots c_{v3,n}$
…
$c_{vn,1}, c_{vn,1}, c_{vn,1}, c_{vn,1} \cdots c_{vn,1}$
to compute CS values in parallel
$c_{1,1}, c_{1,2}, c_{1,3}, c_{1,4} \cdots c_{1,n}$
$c_{2,1}, c_{2,2}, c_{2,3}, c_{2,3} \cdots c_{2,n}$
$c_{3,1}, c_{3,2}, c_{3,3}, c_{3,4} \cdots c_{3,n}$
…
$c_{n,1}, c_{n,1}, c_{n,1}, c_{n,1} \cdots c_{n,1}$
CS matrix C of size n*n
$\begin{bmatrix} c_{1,1} & \cdots & c_{1,n} \\ \vdots & \ddots & \vdots \\ c_{n,1} & \cdots & c_{n,n} \end{bmatrix}$
Parallel computing of MSE values of output word-vectors then MSE values of input word-vectors
to compute MSE values in parallel
$e_{1,1}, e_{1,2}, e_{1,3}, e_{1,4} \cdots e_{1,n}$
$e_{2,1}, e_{2,2}, e_{2,3}, e_{2,3} \cdots e_{2,n}$
$e_{3,1}, e_{3,2}, e_{3,3}, e_{3,4} \cdots e_{3,n}$
…
$e_{n,1}, e_{n,1}, e_{n,1}, e_{n,1} \cdots e_{n,1}$
MSE matrix E of size n*n
$\begin{bmatrix} e_{1,1} & \cdots & e_{1,n} \\ \vdots & \ddots & \vdots \\ e_{n,1} & \cdots & e_{n,n} \end{bmatrix}$

Fig. 6. Computing steps of CPU parallel version of similarity analysis.

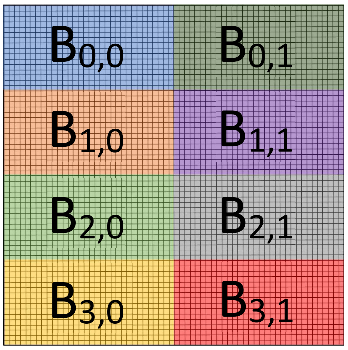


Fig. 7. Possible 8-block $B(i, j)$ division of computed matrices in GPU version of similarity analysis.

TABLE I
COMPUTING TIME VALUES OF SERIAL, CPU PARALLEL AND GPU PARALLEL VERSIONS OF ANALYSIS.

Word Count	Sequential	CPU Parallel	GPU Parallel
1K-word	4.67	8.17	0.5
2K-word	17.66	14.45	0.93
10K-word	307.95	174.79	11.52
20K-word	1098.58	777.61	40.56
30K-word	2146.1	1701.46	86.62
40K-word	3395.67	2732.19	155.94
50K-word	5060.01	4265.16	275.19
60K-word	7026.09	6129.2	576.18
70K-word	9324.22	8276.45	648.48

assessments of the sequential version, the CPU parallel version that utilizes Python's Multiprocessing module and Pool class, and the GPU version using CUDA.

These algorithms are tested on a PC with 16GB RAM; in the GPU parallel version, NVIDIA GeForce GTX 960M (4 GB, 640 cores, GDDR5, 1253 MHz) laptop graphic card was utilized. Additionally, the time spent for device-to-host and host-to-device data transfer was also included in the calculated time costs. Moreover, in the CPU parallel version, Intel Core i7-6700HQ CPU (2.60GHz) was used and each performance test is run with 8 threads (1 thread per core).

In this performance evaluation we used the Blog Authorship corpus, which consists of posts gathered from 19,320 bloggers on blogger.com. The corpus incorporates a total of 681,288 posts and over 140 million words [22].

Due to hardware resource limitations and the nature of the probabilistic similarity analysis algorithm, the serial, CPU parallel and GPU parallel versions of algorithms have reached an upper limit for processed text length that are respectively 73000 words for serial version, 134000 words for the CPU parallel, and 313000 words for the GPU parallel version.

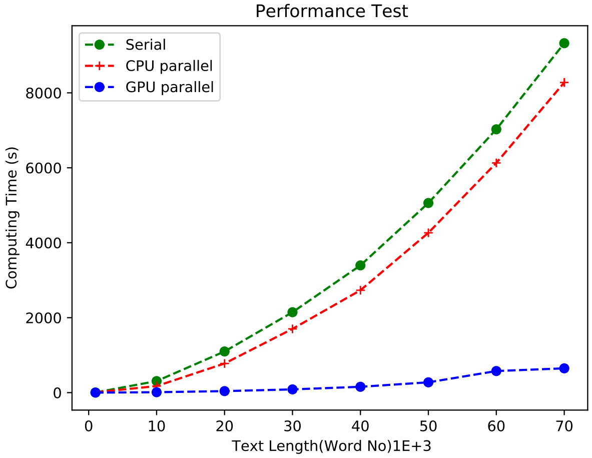


Fig. 8. Computing time of GPU, CPU parallel and serial execution of probabilistic similarity analysis for various text lengths.

Fig. 8 demonstrates the performance comparison of serial, CPU parallel and GPU parallel algorithms by conducting probabilistic similarity analysis. Each algorithm performed analyses on the same text and number of words increased during each iteration by 10K. For this case, the maximum text length is 70K words due to the serial version's upper limit.

The sequential version that serially attains necessary measures and extracts connectivity features of text scores lower computing time than the parallel CPU version for text length less than 2K words. In serial version, computing time then starts increasing drastically for larger data, illustrating quadratic complexity of similarity analysis in accordance with vocabulary set, hence length of text.

The CPU parallel version performs poorly with regard to computing time for text with word number less than 2K since forking data across CPU cores and then joining results of parallel threads creates additional time cost and overheads. This result confirms that using parallel processing and multicore CPUs is not always guaranteed to provide speedup for all sizes of datasets. In order to get benefit of parallel processing, the amount of processed data needs to be increased to monitor speedup and gain performance as shown in Fig. 8.

The GPU version of the analysis shows best performance in Fig. 8 for all text lengths. This outcome is produced by instant computing of multiple data blocks simultaneously without lost computing time for initialization procedures as seen in CPU parallel version. Utilizing GPU cores by CUDA for conducting probabilistic similarity analysis offers minimum processing time in comparison to other versions of the algorithm regardless of data size, hence reducing proportional association between text length and computing time.

Table I shows computing time values that are utilized to create the Fig. 8.

To conclude, Table I shows that the serial version is 1.7 times faster than the CPU parallel version and the GPU parallel CUDA version is 9.3 times faster than the serial version for text length of 1K words. On the other hand, the CPU parallel version is at least 1.2 times (2K words) and at most 1.1 times (70K words) faster than the serial version. The GPU CUDA version is 14.4 times faster than the sequential version. Also, the GPU version is 12.8 times faster than the CPU parallel version for text length of 70K words.

To further explore the results of the CPU parallel and GPU parallel versions, time cost of both versions for text length bigger than 70K words have been processed as shown in Fig. 9. These results are obtained by conducting probabilistic similarity analysis for CPU and GPU parallel versions of the algorithm on the same text starting from 80K to 130K-

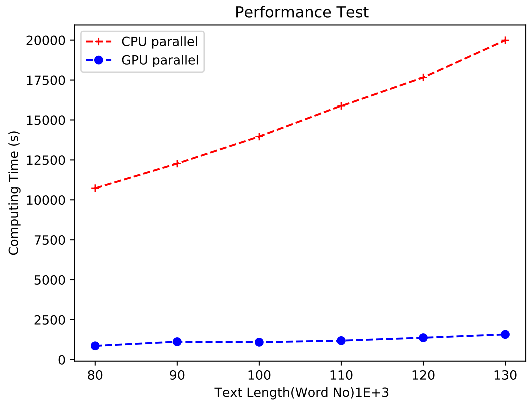


Fig. 9. Computing time of GPU and CPU parallel execution of probabilistic similarity analysis.

TABLE II
COMPUTING TIME VALUES OF SERIAL, CPU PARALLEL AND GPU PARALLEL VERSIONS OF ANALYSIS.

Word Count	CPU Parallel	GPU Parallel
80K-word	10731.88	863.31
90K-word	12273.78	1122.23
100K-word	13963.98	1091.58
110K-word	15881.9	1189.49
120K-word	17660.96	1370.83
130K-word	19990.45	1578.51

word length. Similar to analysis in Fig. 8, each algorithm processed the same data and the number of processed words was increased by 10K words during each iteration.

Table II demonstrates that the GPU parallel version is at least 12.43 times (80K words) and at most 12.66 times (130K words) faster than the CPU parallel version.To sum up, these results show that parallel versions of probabilistic similarity analysis algorithms are promising to utilize applications of similarity based NLP text analysis.

VI. CONCLUSION

In this study, we presented a semantic similarity analysis to indicate relevance of co-related words in text. In linguistics, frequency of co-related lexemes in short text can be useful to utilize probabilistic features of connectivity patterns for semantic elicitation of knowledge conveyed in text. Probabilistic associations of word pairs provide an insight to the textual structure of lexeme sequences. CS and MSE measures can be obtained from input and output-word vectors to denote probabilistic relational similarity of word vector pairs. Moreover, parallel computing is another important aspect in data processing and analytics since the concept of parallel computing has been applied in semantic similarity analysis. In this paper, first, a sequential version is proposed, and then a CPU parallel version is developed and last a GPU parallel CUDA version implemented to get benefits of parallel processing for probabilistic semantics analysis. The results presented in section V indicate that performance limitations of serial similarity analysis are significantly reduced by proposed CPU parallel and GPU parallel versions. Furthermore, this study infers efficiency of utilizing parallel processing techniques and applying graphic processor resources to expand capacity of analyzing probabilistic relations and its indication of similarity analysis among lexemes.

REFERENCES

[1] A. A. Aydin and G. Alaghband, "Sequential and parallel hybrid approach for nonrecursive most significant digit radix sort," in *10th International Conference on Applied Computing*, 2013, pp. 51–58.

[2] S. Berkovich and E. Berkovich, "Method and apparatus for concurrent execution of serial computing instructions using combinatorial architecture for program partitioning," Apr. 8 1997, uS Patent 5,619,680.

[3] A. A. Aydin, "Performance benchmarking of sequential, parallel and hybrid radix sort algorithms and analyzing impact of sub vectors, created on each level,on hybrid msd radix sort's runtime," 2012, mS Thesis, University of Colorado Denver.

[4] B. Parhami, "Parallel processing with big data." 2019.

[5] D. Demirol, R. Das, and D. Hanbay, "Büyük veri üzerine perspektif bir bakış," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, 2019, pp. 1–9.

[6] J. Hromkovič, *Communication complexity and parallel computing*. Springer Science & Business Media, 2013.

[7] A. Aydin and K. Anderson, "Batch to real-time: Incremental data collection & analytics platform," 2017.

[8] S. H. Roosta, "Artificial intelligence and parallel processing," in *Parallel Processing and Parallel Algorithms*. Springer, 2000, pp. 501–534.

[9] T. Strzalkowski, F. Lin, J. Wang, and J. Perez-Carballo, "Evaluating natural language processing techniques in information retrieval," in *Natural language information retrieval*. Springer, 1999, pp. 113–145.

[10] S. Gupta and M. R. Babu, "Performance analysis of gpu compared to single-core and multi-core cpu for natural language applications," *IJACSA Editorial*, 2011.

[11] D. Alnahas and B. B. Alagoz, "Probabilistic relational connectivity analysis of bigram models," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, 2019, pp. 1–6.

[12] I. Dagan, L. Lee, and F. C. Pereira, "Similarity-based models of word cooccurrence probabilities," *Machine learning*, vol. 34, no. 1, pp. 43–69, 1999.

[13] A. M. Schakel and B. J. Wilson, "Measuring word significance using distributed representations of words," *arXiv preprint arXiv:1508.02297*, 2015.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[15] Y. Yin, D. Feng, Z. Shi, and L. Ouyang, "Text recommendation based on time series and multi-label information," 2020.

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[17] S. Zhou, X. Xu, Y. Liu, R. Chang, and Y. Xiao, "Text similarity measurement of semantic cognition based on word vector distance decentralization with clustering analysis," *IEEE Access*, vol. 7, pp. 107 247–107 258, 2019.

[18] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.

[19] R. Raina, A. Madhavan, and A. Y. Ng, "Large-scale deep unsupervised learning using graphics processors," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 873–880.

[20] E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym, "Nvidia tesla: A unified graphics and computing architecture," *IEEE micro*, vol. 28, no. 2, pp. 39–55, 2008.

[21] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with cuda: Is cuda the parallel programming model that application developers have been waiting for?" *Queue*, vol. 6, no. 2, pp. 40–53, 2008.

[22] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, "Effects of age and gender on blogging. aaai spring symposium on computational approaches for analyzing weblogs," 2006.

Dr. Ahmet Arif Aydin is an assistant professor of the Computer Engineering Department at Inonu University. He earned his Ph.D. in Computer Science at the University of Colorado Boulder in 2016 with a specialization in ”Architectural Design for Data Analytics Platforms”. His current research interests include software engineering, data-intensive system design, crisis informatics, big data analytics, data modeling, algorithm design for analytics, parallel processing, and machine learning.

Dima Alnahas currently works as a Software Development and AI Team Leader at Infina Software Inc. She is pursuing her Masters’ Degree at Computer Engineering Department in Kadir Has University. She also attends Baden-Wuerttemberg Cooperative State University (DHBW) as an exchange masters’ student with research interests in Natural Language Processing and Machine Learning.

Publication Ethics

The journal publishes original papers in the extensive field of Electrical-electronics and Computer engineering. To that end, it is essential that all who participate in producing the journal conduct themselves as authors, reviewers, editors, and publishers in accord with the highest level of professional ethics and standards. Plagiarism or self-plagiarism constitutes unethical scientific behavior and is never acceptable.

By submitting a manuscript to this journal, each author explicitly confirms that the manuscript meets the highest ethical standards for authors and coauthors

The undersigned hereby assign(s) to *Balkan Journal of Electrical & Computer Engineering* (BAJECE) copyright ownership in the above Paper, effective if and when the Paper is accepted for publication by BAJECE and to the extent transferable under applicable national law. This assignment gives BAJECE the right to register copyright to the Paper in its name as claimant and to publish the Paper in any print or electronic medium.

Authors, or their employers in the case of works made for hire, retain the following rights:

1. All proprietary rights other than copyright, including patent rights.
2. The right to make and distribute copies of the Paper for internal purposes.
3. The right to use the material for lecture or classroom purposes.
4. The right to prepare derivative publications based on the Paper, including books or book chapters, journal papers, and magazine articles, provided that publication of a derivative work occurs subsequent to the official date of publication by BAJECE.
5. The right to post an author-prepared version or an of ficial version (preferred version) of the published paper on an i nternal or external server controlled exclusively by t he author/employer, pr ovided that (a) such posting is noncommercial in nature and the paper is made available to users without charge; (b) a copyright notice and full citation appear with the paper, and (c) a link to BAJECE’s official online version of the abstract is provided using the DOI (Document Object Identifier) link.

bajece

BAJECE
Balkan Journal of
Electrical & Computer
Engineering
An International Peer Reviewed, Indexed and Open Access Journal

ISSN: 2147- 284X
Year: October 2022
Volume: 10
Issue: 4

CONTENTS

BALKAN JOURNAL OF
ELECTRICAL & COMPUTER ENGINEERING
(An International Peer Reviewed, Indexed and Open Access Journal**)**

Contact
Batman University
Department of Electrical-Electronics Engineering
Bati Raman Campus Batman-Turkey

Web: http://dergipark.gov.tr/bajece
http://www.bajece.com
e-mail: bajece@hotmail.com