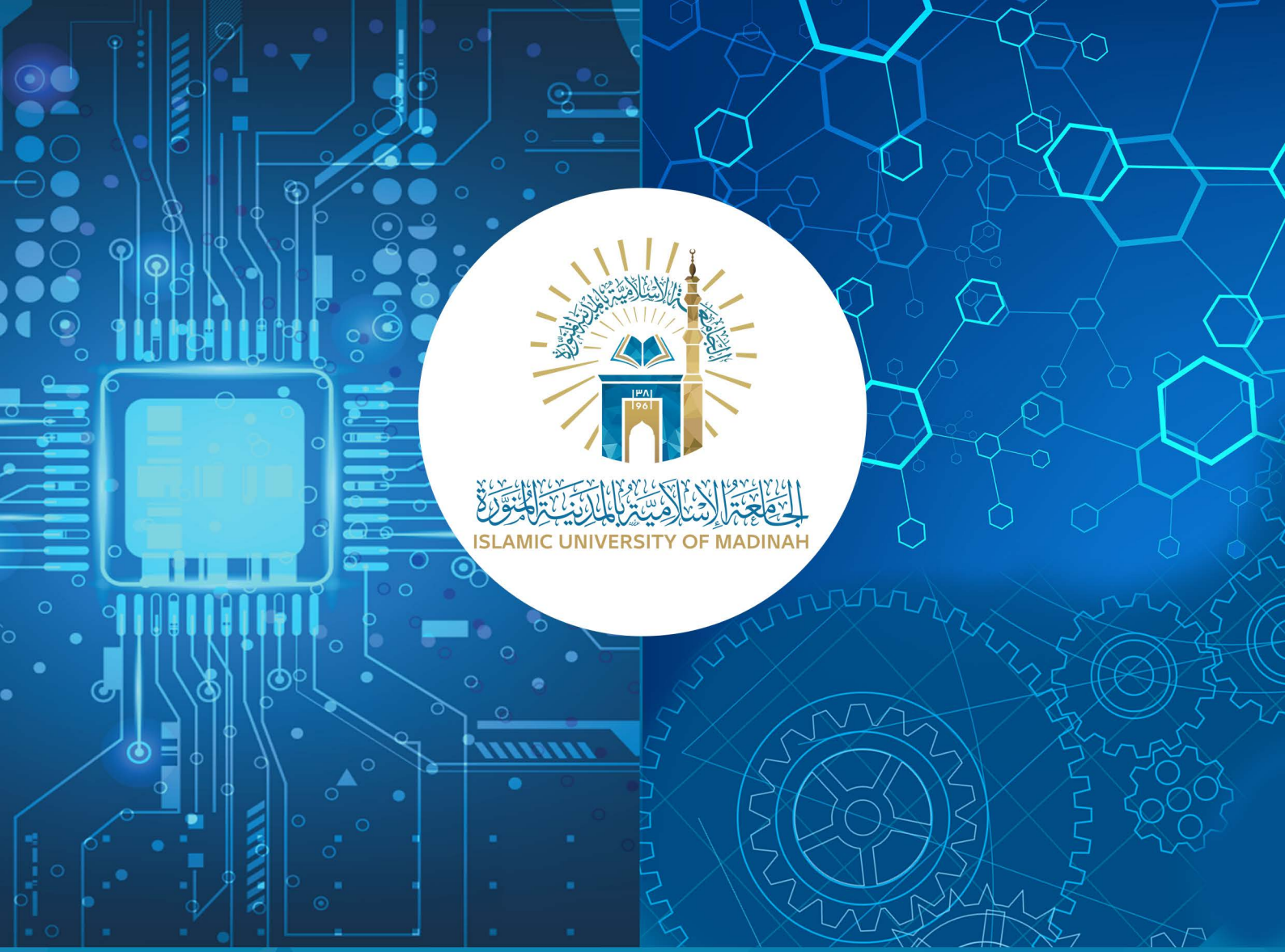


Journal of

Engineering, Science & Computing

Vol: 2 Issue: 2 Year: 2020


بسم الله الرحمن الرحيم

Paper version

Filed at the King Fahd National Library No. 8742/1439 on 17/09/1439 AH

International serial number of periodicals (ISSN) 1658-7936

Online version

Filed at the King Fahd National Library No. 8742/1439 on 17/09/1439 AH

International Serial Number of Periodicals (e-ISSN) 1658-7944

The Journal's Website

https://jesc.iu.edu.sa

The papers are sent in the name of the Editor-in-Chief of the Journal to this E-mail address

jesc@iu.edu.sa

Publication Rules at the Journal (*)

❖ **General rules:**

- Report original scientific research (the main results and conclusions must not have been published or submitted elsewhere).
- Fit with the topics of the journal.
- Report novel results, innovative work and show a new scientific contribution.
- Not to bear similarity of more than 25% of a previously published work of the same author(s).
- Follow the rules, regulation and authentic research methodologies.
- Fulfill the required items and the format of the journal provided in appendix below related to the guide for author.
- Opinions expressed in published articles commit the authors themselves only and not necessarily the opinion of the journal.

❖ **For all articles:**

- The exclusive right to publish and distribute and article, and to grant rights to others, including commercial purposes.
- For open access articles, IU will apply the relevant third-party user license where IU publishes the article on its online platforms.
- The right to provide the article in all forms and media so the article can be used on the latest technology even after publication.
- The authority to enforce the rights in the article, on behalf of an author, against third parties, for example in the case of plagiarism or copyright infringement.

(*) These general rules are explained in details along with other rules for Author's guide in the journal's website: https://jesc.iu.edu.s

(JESC)

The Journal of Engineering, Science and Computing

Issued By

Islamic University of Madinah, Saudi Arabia

Dear friends and colleagues,

With great pleasure, I am delighted to present you the fourth issue of the Journal of Engineering, Science, and Computing (JESC) managed and edited by the Islamic University of Madinah, Saudi Arabia.

The mission of the JESC is to disseminate contemporary research work from researchers across the globe in the areas of Engineering, Science, and Computing.

The fourth issue of the journal attracted a significant number of submissions, which brought great pleasure to us. However, our policy stays put, and we pledged to present to you quality research. Therefore, this issue comes with ten articles which represent an acceptance rate of around 40%. The editorial board including the Editor-in-Chief and the editorial advisory board would like to thank all of those who contributed to this issue and would like to welcome potential authors to contribute to future issues.

It is a privilege to share with you the fourth issue of the journal and wish you happy reading.

Mohammad Abdeen, Ph.D., P.Eng.

Editor-in-Chief

The Journal of Engineering, Science and Computing (JESC)

Table of Content

Recent Developments in Year 2000s in the Modelling of Turbulence Transport in Catalysis

Oyegoke Toyese,
Laboratoire de Chimie, ENS, l'Universite de Lyon, Lyon, France
Department of Chemical Engineering, Ahmadu Bello University Zaria, Nigeria
Toyese.oyegoke@ens-lyon.fr, OyegokeToyese@gmail.com
and
Adamu Yusuf,
Department of Chemical Engineering, Ahmadu Bello University Zaria, Nigeria
Edomwonyi-Otu C. Lawrence,
Department of Chemical Engineering, Ahmadu Bello University Zaria, Nigeria.
Department of Chemical and Petroleum Engineering, Delta State University, Nigeria
Akanji N. Abiodun
Department of Chemical Engineering, Ahmadu Bello University Zaria, Nigeria.
Department of Chemical Engineering Technology, Federal Polytechnic Bida, Nigeria

Abstract: The turbulent nature of catalytic reactions has been well reported. For some reactions, the higher the rate of turbulence, the faster the reaction process. This paper focus on the review of various research works where turbulence models were employed in promoting and advancing study and knowledge of catalysis or catalytic reaction systems (such as fixed bed reactor, trickle bed reactor, combustor, among others) or processes in the twentieth centuries. It also draws attention to several fluid computational dynamics package employed in the simulation and different contributions that have been made in advancing research in the field of catalysis via turbulence modeling. The essence of these is to enhance effective and efficient reactant access to the active sites of the catalyst. This study, however, shows that models such as k–e and RSM turbulence models are better suited for predicting or studying turbulence behavior in a catalytic reaction. It was realized that apart from selecting the turbulence model, appropriate selection of the kinetic model plays a significant role in promoting accurate prediction when carrying out simulations. However, this study was able to identify that only a few research works have given attention to the right and appropriate use or selection of a kinetic model for catalytic reaction systems.

Keywords: Simulation, Transport Phenomena, Kinetic Model, Chemical Reactors, Turbulence.

التطورات الأخيرة في العام ٢٠٠٠ في نمذجة النقل المضطرب بالتحفيز

الملخص: تم الإبلاغ جيدًا عن الطبيعة المضطربة للتفاعلات التحفيزية. بالنسبة لبعض التفاعلات، كلما زاد معدل الاضطراب، زادت سرعة عملية التفاعل. تركز هذه الورقة على مراجعة الأعمال البحثية المختلفة حيث تم استخدام نماذج الاضطراب في تعزيز وتطوير الدراسة والمعرفة بأنظمة الحفز أو التفاعل التحفيزي (مثل مفاعل السرير الثابت، مفاعل السرير المقطر، جهاز الاحتراق، من بين أمور أخرى) أو العمليات في القرن العشرين. كما يلفت الانتباه إلى العديد من حزم ديناميكيات حساب السوائل المستخدمة في المحاكاة والمساهمات المختلفة التي تم إجراؤها في تطوير البحث في مجال التحفيز عبر نمذجة الاضطراب. يتمثل جوهر هذه العناصر في تعزيز الوصول الفعال والكفؤ للمتفاعلات إلى المواقع النشطة للمحفز. ومع ذلك، تُظهر هذه الدراسة أن نماذج مثل نماذج الاضطراب k-e وRSM هي أكثر ملاءمة للتنبؤ أو دراسة سلوك الاضطراب في تفاعل تحفيزي. تم إدراك أنه بصرف النظر عن اختيار نموذج الاضطراب، يلعب الاختيار المناسب للنموذج الحركي دورًا مهمًا في تعزيز التنبؤ الدقيق عند إجراء عمليات المحاكاة. ومع ذلك، تمكنت هذه الدراسة من تحديد عدد قليل فقط من الأعمال البحثية التي قد أعطت اهتمامًا للاستخدام الصحيح والمناسب أو اختيار نموذج حركي لأنظمة التفاعل التحفيزي.

1. Introduction

Turbulence flow is a type of flow regime in fluid dynamics characterized by flow velocity and disordered change in pressure [1]. This phenomenon is often observed in routine processes like billowing storm clouds, fast-flowing river, surf, or smoke from its different bases. Turbulent Fluid flows occur in the natural and engineered system and other related applications [2-4]. It is commonly driven by substantial kinetic energy within the flowing fluid in excess of the prevailing viscous absorbing effect of the flows. This shows that turbulent effects are more challenging to create in high viscosity fluids than in low viscous fluids. The prediction of turbulence is often quantified by a dimensionless physical quantity (or constant) known as Reynolds number from inception, which signifies the proportion of kinetic energy to the amount of the viscous damping present in a flow. However, due to the lengthy resistant detailed physical analysis in turbulence coupled with the interactions holding within, which created a very complicated situation. Richard Feynman has identified that turbulence as a crucial unresolved challenge in the classical physics field [5].

Turbulence is often identified to be irregularity or randomization, which indicates that turbulence problems can be solved statistically instead of solving the problem deterministically. Also, it is known to be chaotic and disorganized. However, not all disorganized flows are turbulent [6]. Turbulence flows are also characterized by diffusivity (which tends to increase the homogeneity of the mixing components in the fluid), rotationality, and dissipation. The turbulent flows which are dissipative make it necessary to make provision for a constant generation of energy supply to sustain it like in the case of hydro-power supply. The dissipation of turbulence does hold when a substantial kinetic energy present in the fluid gets converted into internal energy employing viscous shear stress [7].

This paper focus on the review of different research works where turbulence models are employed in promoting and advancing studies and knowledge of catalysis or catalytic reaction systems or processes in the twentieth century.

2. Catalysis

Catalysis, a field of study that deals with reaction systems and its ways of speeding up the number of chemical species interaction and conversion due to the influence of a catalyst. In this process, the catalyst is not consumed but can be repeatedly used. During these chemical reaction processes, catalysts are not used up or consumed by the primary processes but can be deactivated, prevented, or destroyed by secondary processes (that is, side reactions). For instance, side reactions like dissolution, coking, scission, or sublimation, on a heterogeneous

catalyst [8-11]. In most cases, only a small quantity of catalysts is required in principle. In general, the catalyst influences the reactions to hold faster due to lower energy requirements due to the low activation energy [11, 12].

Studies have revealed that catalyzed reactions display lower activation energies than the corresponding analyzed un-catalyzed reactions, yielding a higher rate of chemical reaction, at the same thermal conductivity and the same reaction density. However, the details of the mechanics for catalysis are intricate. Catalysts fundamentally influence the environment of the reaction and support the reagents to have the positive bonds polarized, e.g., Acid catalysts used in the formation of naturally occurring intermediates (specific), or the reactions of carbonyl compounds, like osmate esters present in the dehydroxylation of osmium tetroxide-alkynes. Physically, most catalytic reactions are often chemical reactions; The rate of these chemical reactions often depends on the reaction contact frequency with the catalyst surface when overcoming the energy complex of the rate-determining step (RDS) in its reaction path [13]. Generally, the catalyst's participation is at a slower pace, and the rates are controlled via the amount and activity of the catalyst. The diffusion rate to the reaction surface and the diffusion behavior of the products from the surface can be easily determined in heterogeneous catalysis. ZSM-5 is an excellent example of a solid catalyst [14].

Catalysts may be categorized into either heterogeneous (otherwise known as a solid catalyst) or homogeneous catalysts, and other bio-catalysts, otherwise known as enzymes, are often considered a third category. Several works in diverse areas of research have employed different catalysts to enhance their reaction processes or steps. Some of these works that used homogeneous catalysts are the work of Vicente et al. [15] in biodiesel production where nearly 100 wt% yield was obtained for the use of methoxide catalyst while the review of Johnson et al. [16] report indicates the use of the homogenous catalyst for the dehydrogenation of alcohol for the production of hydrogen will only be feasible low molecular weight alcohols (MWA) like methanol, ethanol, and isopropanol, unlike larger MWA.

Whereas those that employ the use of heterogeneous catalyst includes Helwani et al. [17] review unveiled that the use of a solid catalyst better enhances the triglycerides transesterification process than enzymes; Oyegoke et al. [18] in the dehydrogenation of propane into propylene employed the use of solid catalyst where the studies indicate that the Cr site was the most active and reactive site the promote the hydrogen abstraction from the feed. Further, Zhao et al. [19] in solar water oxidation showed a high activity towards the reaction, while Morales-Delarosa et al. [20] in the cellulose hydrolysis into fermentable sugar,

and many others with a resulting conversion >99 % after 5 h of reaction and with 87 % glucose yield.

Biocatalysts are developed by Gao et al. [21] for the enhancement of enzymatic reaction, which results in glycerol conversion and Glycerol carbonate yield as 85.20 %, and 64.71%, respectively after 24 h. Likewise, Cubides-Roman et al. [22] employed the use of an enzyme to accelerate biodiesel production, resulting to a low optimum conversion 9.6 %, while Das et al. [23] studies indicate that the prepared nano biocatalysts appear to be potent catalysts for successful industrial applications, especially in food industries. In general, the catalyst could be some transition metals, transition metal complexes, and/or enzymes.

3. Mechanism of Catalytic Reactions

According to Matthiesen et al. [24], catalysts often interact with different reacting species in a chemical reaction to form intermediate surface species (which are often said to be unstable product) that subsequently gets transformed into the desired products and some undesired ones. The presence of theses undesired products that often get stocked on the catalyst surface has made it necessary to regenerate the catalyst by getting rid of the deposited materials' surface before recycling it. The scheme presented in equations 1 to 5 is a reaction system that is common in which C stands for the catalyst, with P as a product while A and B represent reacting species:

$$C + A \rightarrow CA \quad (1)$$

$$B + CA \rightarrow BCA \quad (2)$$

$$BCA \rightarrow CP \quad (3)$$

$$CP \rightarrow C + P \quad (4)$$

The scheme indicated that the catalyst (C) was used up by reaction 1, and it was after that released in reaction 4, so the catalyst is not always displayed in the chemical expression for the overall reaction in equation 5, indicating that catalysts are neither transformed nor destroyed [25, 26].

$$A + B \rightarrow P \quad (5)$$

The recovery of catalysts shown in the scheme indicates the only a small amount is needed to influence the speed of a chemical reaction. Sometimes, catalysts are used-up in secondary processes, not often. Catalyst is not expressed in a rate expression (or equation), e.g., if the foremost step in the displayed reaction process is taken as the one that determines the reaction rate (rate-determining step, RDS), the network of catalyst influenced reaction is then a second-order reaction with the rate given as; $v = k_{cat}[A][C]$ when RDS approximation is adopted for the rate, which indicates that it varies directly with the catalyst concentration [C]. The [C]

remains unchanged for the period of the reaction, while the catalyzed reaction is pseudo-first-order written in the form: $v = k_{obs}[A]$, where $k_{obs} = k_{cat}[C]$ (Matthiesen et al. [24]). A catalyst can influence the change in equilibrium concentration in a continuous process, but this is, however, contrary to the laws of thermodynamics [24].

4. Turbulence Models Used in Catalysis

Modeling of turbulence is the method of constructing and making use of models (in forms of mathematical expressions) for the prediction and comprehension of turbulence phenomena in engineering or scientific processes or systems. Turbulent fluid flow exhibits traits or capabilities on different scales like timescale, length-scale, and so forth, which all interact with each other [27]. In the work of Ching [28], it was identified that a common approach employed in the field of turbulence modeling is to norm the main equations of the flow rate and to zero in on large-scale and flow's laminar components.

Nevertheless, it is necessary to model the effects of the fluctuating and small scales region in turbulence [28, 27]. A further survey of the literature unveils that several turbulence transport-models such as K-epsilon (k–ε), Reynolds Stress Model (RSM), and Spalart Allmaras (*s-a*) have been employed in the simulation of computational fluid dynamics (CFD). Table 1 displayed the various CFD based research works done so far. Further reports in this section present a summary of commonly applied turbulence equations and models that have been found useful in recent reaction engineering processes, and reports of other research work where they are employed with special attention for catalysis-based studies.

4.1 k–epsilon (k–ε)

The typical model often applied in CFD is the K-epsilon (k-ε) turbulence equations, which could be used to study mean flow properties for turbulent flow situations. It is can also be known as a two-equation model that reveals the general behavior of turbulence via the use of two (2) transport models (PDEs). As reported by [28], the kinetic energy of turbulence, k, can be expressed as:

$$\frac{\partial(\rho k)}{\partial t} + \frac{\partial(\rho k u_i)}{\partial x_i} = \frac{\partial}{\partial x_j}\left[\frac{\mu_t}{\sigma_k}\frac{\partial k}{\partial x_j}\right] + 2\mu_t E_{ij}E_{ij} - \rho\varepsilon \tag{6}$$

While the dissipation [28], ε will be in the form:

$$\frac{\partial(\rho \varepsilon)}{\partial t} + \frac{\partial(\rho \varepsilon u_i)}{\partial x_i} = \frac{\partial}{\partial x_j}\left[\frac{\mu_t}{\sigma_\varepsilon}\frac{\partial \varepsilon}{\partial x_j}\right] + C_{1\varepsilon}\frac{\varepsilon}{k}2\mu_t E_{ij}E_{ij} - C_{2\varepsilon}\rho\frac{\varepsilon^2}{k} \tag{7}$$

Where u_i represents velocity component in the corresponding direction, E_{ij} represents the component of the rate of deformation, μ_t represents eddy viscosity given as:

$$\mu_t = \rho C_\mu \frac{k^2}{\varepsilon}$$

The models also consist of some adjustable constants such as $\sigma_k, \sigma_\varepsilon, C_{1\varepsilon}$, and $C_{2\varepsilon}$. These constant values have were obtained via numerous iterations of data fitting for a wide range of turbulent flows which were reported as $C_\mu = 0.09$, $\sigma_k = 1.00$, $\sigma_\varepsilon = 1.30$, $C_{1\varepsilon} = 1.44$, and $C_{2\varepsilon} = 1.92$ in literature.

The improved model of the mixing-length equations was identified to be the K-epsilon model, including the provision of a replacement to the algebraically prescribed turbulence and fluctuating length-scale of measurement in medium to very complex flows. Research works of Christoph et al. [29] employed the *k-e* representative equations to investigate turbulence in a combustion process, within channel flows, of the mixture of hydrogen and air over a stabilized platinum catalyst. With the use of a *k-e* model, Muhammad [30]) studied the CFD prognostication of profile displaying mass fraction of gasoline and gas oil in a fluid catalytic cracking (FCC) riser main. Using the k-epsilon turbulence model and 4-lump model, the same author in 2013 also examined an FCC riser to predict gasoline yield.

In 2014, an analysis of the material flow within a radial flow fixed bed (FB) reactor was conducted by Dominick et al. [31]. Following this, Afshin [32] evaluated the effect of carbon tetrachloride (CCl_4) addition as propulsion to the thermal cracking reactor as a result of the coke formed in separate coil outlet temperatures (COT). Both Dominick et al. [31] and Afshin [32] employed *k-ε* models.

This model is simple, affordable, and considerably accurate for a wide variety of flow that lacks separation. Literature indicates that this model often failed to give an accurate prediction for the onset and the flow separation under the adverse pressure gradient [1, 28, 33].

4.2 Menter's Shear Stress Transport (SST)

SST (Menter's shear stress transport model) turbulence model is one of the most robust two-equation eddy-viscosity turbulence models often used in CFD [33]. The model combines both the k-omega turbulence model and the K-epsilon turbulence model, such that the k-omega model is employed in the boundary layer inner region and switches to the k-epsilon model in the free shear flow system.

$$\frac{\partial(\rho k)}{\partial t}+\frac{\partial(\rho u_j k)}{\partial x_j}=P-\beta^*\rho\omega k+\frac{\partial}{\partial x_j}\left[\left(\mu+\sigma_k\mu_t\right)\frac{\partial k}{\partial x_j}\right] \tag{8}$$

$$\frac{\partial(\rho\omega)}{\partial t}+\frac{\partial(\rho u_j \omega)}{\partial x_j}=\frac{\gamma}{\nu_t}P-\beta\rho\omega^2+\frac{\partial}{\partial x_j}\left[\left(\mu+\sigma_\omega\mu_t\right)\frac{\partial\omega}{\partial x_j}\right]+2(1-F_1)\frac{\rho\sigma_{\omega 2}}{\omega}\frac{\partial k}{\partial x_j}\frac{\partial\omega}{\partial x_j}$$

(9)

Here, $P=\tau_{ij}\frac{\partial u_i}{\partial x_j}$ $\tau_{ij}=\mu_t\left(2S_{ij}-\frac{2}{3}\frac{\partial u_k}{\partial x_k}\delta_{ij}\right)-\frac{2}{3}\rho k\delta_{ij}$,

$S_{ij}=\frac{1}{2}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)$, $\mu_t=\frac{\rho a_1 k}{\max(a_1\omega,\Omega F_2)}$,

$\phi=F_1\phi_1+(1-F_1)\phi_2$, $F_1=\tanh\left(\arg_1^4\right)$, $\arg_1=\min\left[\max\left(\frac{\sqrt{k}}{\beta^*\omega d},\frac{500\nu}{d^2\omega}\right),\frac{4\rho\sigma_{\omega 2}k}{\mathrm{CD}_{k\omega}d^2}\right]$,

$\mathrm{CD}_{k\omega}=\max\left(2\rho\sigma_{\omega 2}\frac{1}{\omega}\frac{\partial k}{\partial x_j}\frac{\partial\omega}{\partial x_j},10^{-20}\right)$, $F_2=\tanh\left(\arg_2^2\right)$, $\arg_2=\max\left(2\frac{\sqrt{k}}{\beta^*\omega d},\frac{500\nu}{d^2\omega}\right)$

The constants $\sigma_{k1}=0.85,\sigma_{w1}=0.65,\beta_1=0.075$ for the k-w closure, the constants for the k-e closure, where $\sigma_{k2}=1.00,\sigma_{w2}=0.856,\beta_2=0.0828$, while $\beta^*=0.09,a_1=0.31$ were the SST's closure constants. The boundary/wall/far-field conditions are given as follows:

$$\frac{U_\infty}{L}<w_{\text{farfield}}<10\frac{U_\infty}{L},\frac{10^{-5}U_\infty^2}{Re_L}<k_{\text{farfield}}<\frac{0.1U_\infty^2}{Re_L},\ \omega_{wall}=10\frac{6\nu}{\beta_1(\Delta d_1)^2},k_{wall}=0$$

The SST model has reported being highly accurate for the predictions of the onset and the amount of flow separation under adverse pressure gradients via the introduction of transport effects into the formulation of the eddy-viscosity. Also, despite being a blend of k-ω and k-ε, that is a k-ω model near the wall and transitions to a k-ε model away from the wall. The model has been reported to be relatively insensitive to the free stream value of ω. [1, 28, 33]

4.3 k–omega (k–ω)

In CFD, the k–omega (*k–ω*) turbulence representative equation is a popular two-equation model, which is often employed as a solution for the Reynolds-averaged Navier–Stokes (RANS) equations [34]. The representative equation is designed to unveil turbulence profiles via the use of 2-partial derivative equations for 2-variable quantity k and ω, where k is the upheaval kinetic energy. In contrast, ω is the specific dissipation rate of the kinetic energy (k) of turbulence to obtain internal thermal energy.

Mathematically, the eddy viscosity ν_T, as required in the RANS equations, is expressed as $\nu_T = k/\omega$, and the development of k and ω is mathematically represented [28] in the form:

$$\frac{\partial(\rho k)}{\partial t}+\frac{\partial(\rho u_j k)}{\partial x_j}=\rho P-\beta^*\rho\omega k+\frac{\partial}{\partial x_j}\left[\left(\mu+\sigma_k\frac{\rho k}{\omega}\right)\frac{\partial k}{\partial x_j}\right],\qquad \text{with } P=\tau_{ij}\frac{\partial u_i}{\partial x_j}, \tag{10}$$

$$\frac{\partial(\rho\omega)}{\partial t}+\frac{\partial(\rho u_j \omega)}{\partial x_j}=\frac{\alpha\omega}{k}P-\beta\rho\omega^2+\frac{\partial}{\partial x_j}\left[\left(\mu+\sigma_\omega\frac{\rho k}{\omega}\right)\frac{\partial\omega}{\partial x_j}\right]+\frac{\rho\sigma_d}{\omega}\frac{\partial k}{\partial x_j}\frac{\partial\omega}{\partial x_j} \tag{11}$$

Where k is the turbulence kinetic energy, ω is the specific rate of dissipation (of the turbulence kinetic energy k into internal thermal energy), ν_T is kinematic eddy viscosity. The coefficients were reported as $\alpha=5/9$, $\beta=3/40$, $\beta^*=9/100$, $\sigma=1/2$, $\sigma^*=1/2$, and $\varepsilon=\beta^*\omega k$.

This model has been reported to have shown an advantageous feature for integrating through a viscous sub-layer and inflows with adverse pressure gradients, being a two-equation model, which have adequate representation for both scales independently. The literature further indicates that the k-ω models are more accurate and numerically stable in the near-wall region [1, 28, 34].

4.4 Reynolds Stress Model (RSM)

The RSM, also known as the second-moment closure representative equations, is the most complete of the classical turbulence modeling approach. It is mathematically expressed as follows:

$$\frac{DR_{ij}}{Dt}=D_{ij}+P_{ij}+\Pi_{ij}+\Omega_{ij}-\varepsilon_{ij} \tag{12}$$

Where R_{ij} is $\langle u_i' u_j' \rangle = -\tau_{ij}/\rho$, D is diffusion-controlled transport of R_{ij}, P is R_{ij} production speed, Π is R_{ij} transport resulting from interactions of turbulent pressure-strain, Ω is R_{ij} transport resulting from rotation, ε is dissipation speed of R_{ij}.

The eddy-viscosity representative equations are well known and include the *k–e* (k–epsilon) and the *k–ω* (k–omega) equations, which have weaknesses in sophisticated flows applicable in engineering. The findings were a result of the use of hypothesizing with the eddy-viscosity equations in their articulation. For example, in a flow with high levels of anisotropy, separation of flows, areas of recirculating flow, rotational-effects flows, or the significant streamline curvature, these models performed very poorly [35]. In these sets of flows, RSM offers superior prediction quality [36]. Closures based on Eddy viscosity are not able to explain the reversal to the isotropy of turbulence, which is noticed in the decaying of turbulent flows [37]. Models based on Eddy viscosity are not able to cannot produce replicas of the properties of a turbulent

flow in the fast-paced deformation limit, where a flexible system indicates the behavior of a turbulence flow [38].

Reynolds stress models have been reported to have displayed better accuracy significantly when compared with turbulence models although, it is expensive (that is, significantly slower to solve its set of equations), and it requires a good initial guess. Generally, suffer from numerical stability issues due to the complexity of the modeled terms, which was one of the key reasons why the models are not regularly employed in a large-scale problem [33, 34].

Table 1: Different applications of turbulence models in catalysis

Author	Research Works	Turbulence Model (s) Used
Christoph et al. [29]	An experimental and numerical study of turbulent in a channel flow combustion (CST) of air-hydrogen mixtures over stabilized platinum catalysts.	Closure of Turbulence was attained via Low Reynold *k-ε* models.
Rodrigo and Rosa [40]	Modeling of turbulence present in a multi-phase flow in a high-pressure trickle-bed reactor (TBR).	*S-k-ε*, *R-k-ε*, RNG and RSM models.
Binxin [41]	Examination of turbulence models for non-Newtonian fluid flow in anaerobic digesters via the use of the CFD approach.	low-Reynolds-number (LRe) k-ε, high-Reynolds-number (HRe) *k-ε, k-ω*, and the RSM.
Muhammad [30]	The prediction of gas oil and gasoline's mass fraction profiles in an FCC riser via the CFD approach.	*k-ε* model
Muhammad [43]	Using CFD to predict the yield of gasoline in an FCC riser by applying *k*-epsilon turbulence and 4-lump kinetic representative equations.	*k-ε* model
Xiaomin et al. [39]	Using CFD modeling for the study of reaction kinetics in a catalytic dehydrogenation of syngas in a fixed-bed reactor	Spalart–Allmaras *(s-a*)
Dominick et al. [31]	The flow was analyzed in a radial flow fixed bed reactor	A RANS type *k-ε* turbulence model
Zhapbasbayev et al. [42]	Modeling of turbulent flow behavior in a radial reactor with a fixed bed configuration	Reynolds motion equations and k-ε model
Afshin [32]	The addition of CCl_4 propulsion to the thermal cracking reactor as a result of the coke produced in separate coil outlet temperatures (COT) was evaluated	Standard *k-ε*, and RNG *k-ε* models.

4.5 Spalart–Allmaras (s–a)

One of the one-equation models that solve a modeled transport mathematical expression for motion-focused eddy fluctuating viscosity is the Spalart–Allmaras (*s-a*) model. The model was explicitly developed for systems that involve flows in the direction of the wall and showed that good results were obtained when boundary layers were controlled by opposing pressure gradients. Progressively, it is being accepted as essential in turbo-machinery applications.

The model is expressed as follows:

$$\frac{\partial \tilde{\nu}}{\partial t} + U_j \frac{\partial \tilde{\nu}}{\partial x_j} = P_{\tilde{\nu}} - \epsilon_{\tilde{\nu}} + \frac{\partial}{\partial x_j}\left[\frac{1}{\rho}\left(\mu + \frac{\tilde{\nu}}{\sigma_{\tilde{\nu}}}\right)\frac{\partial \tilde{\nu}}{\partial x_j}\right] \quad (13)$$

A survey of literature has indicated that this model does not accurately compute fields that exhibit separated flow, shear flow, or decaying turbulence due to the absence of a correction factor for compressibility, resulting in the model's overpredict the growth rate of high-speed shear layers. Although, the model has shown for giving good results for boundary layers and adverse pressure gradients [1, 28, 34].

Moreover, Xiaomin et al. [39] modeled the turbulence behavior of the reacting species in the use of a catalyst to remove hydrogen from syngas in a fixed-bed (FB) reactor by applying the *s-a* model to unveil some information about the reaction kinetics.

Some other research works employed several turbulence models to screen out the best model that gave the best prediction for the concern system chosen for the study. In line with this approach, Rodrigo and Rosa [40] employed the use of *k-e* and RSM models for the representation of turbulence behavior in two or more fluid flows in trickle bed reactors under high pressure. At the same time, Binxin [41] investigated the turbulence representative equations for non-Newtonian fluid flow behavior in an anaerobic digester using the *k-ε*, RSM, and *k-ω* representative equations.

Furthermore, Zhapbasbayev et al. [42] model a turbulent flow present in a radial reactor with a fixed bed using Reynolds motion equations and *k-ε* models. *S-k-ε*, *R-k-ε,* RNG, RSM, and *s-a* can be used for a fixed bed reactor to model turbulence transport in catalysis. However, the review shows that *k-e* models are commonly employed in this practice, unlike RNG, RSM, or *s-a*.

5. CFD Packages Commonly Used in the Modelling of Turbulence in Catalysis

Computational fluid dynamics with the acronym 'CFD' is a field of study in fluid mechanics that deals with the use of numerical analysis and data structures to provide solutions and analytical parameters for problems that involve fluid flow. In this area of fluid mechanics, computers are employed to execute several calculations or computations required to simulate or evaluating the interaction of gases and liquids with surfaces distinction by boundary conditions. The use of CFD analysis has made it realistic to enable quick, efficient simulation of heat transfer and fluid flow through the use of software (Solid Works). Examples of the software or packages are Gambit, FLUENT, COMSOL Multi-physics, and many others. These

applications have long been employed in the study of the turbulence flow in a catalytic system such as combustors, TBRs, bio-digesters, FCC riser reactor, fixed-bed reactors, cracking reactors, and so on as presented in the literature [29-32, 39-42].

Table 2 - Computational fluid dynamic packages commonly used in catalytic systems

Author	System Studied	Software Used
Christoph et al. [29]	Hydrogen/air mixtures in a bed of platinum in a combustor.	CHEMKIN
Rodrigo and Rosa [40]	Flow in trickle bed reactor (TRB)	Fluent 6.1 software, Gambit, PC-SIMPLE.
Binxin [41]	Studied flow in an anaerobic digester	Gambit 2.4.6 and Fluent 12.0
Muhammad [30]	FCC riser reactor (Alumina used as a catalyst)	FLUENT 6.3 was employed.
Muhammad [43]	FCC riser reactor (Alumina used as a catalyst)	FLUENT 6.3
Xiaomin *et al.* [39]	Syngas in a heterogeneous fixed-bed reactor (FBR)	FLUENT 6.3.26, SIMPLE
Dominick et al. [31]	Evaluate of radial flow behavior in a fixed bed reactor	COMSOL Multi-physics
Zhapbasbayev*et al* [42]	Flow in a duct with porous medium (or flow in a pipe with a fixed bed)	Not reported
Afshin [32]	Thermal cracking reactor	Not reported

A review of the literature shows that the FLUENT package has been the most common package employed in CFD analysis of the catalysis, turbulence modeling works like Rodrigo and Rosa [28], Binxin [41], Muhammed [30, 43], and Xiaomin et al. [39] were some of the authors that employed the use of FLUENT package for their studies. Table 2 summarizes different CFD packages commonly used in catalysis. The study of works that employed the use of CFD packages to analyze their systems is Christoph et al. [29] that CHEMKIN used for chemical kinetics and laminar transport data evaluation of hydrogen-air combustor. Moreover, Binxin [41] employed the used Gambit 2.4.6 for meshing, while Fluent 12.0 was used to solving the governing equations.

However, Muhammad [30], who studied the reactor (FCC riser) using alumina as a catalyst, employed the used FLUENT 6.3 for discretization, meshing, and governing equations solution. A similar approach was employed by Xiaomin et al. [39], unlike Rodrigo and Rosa [40], who employed the use of Fluent 6.1, Gambit, and PC-Simple, to obtained accurate results.

6. Turbulence Modelling as a Means of Enhancing Studies in Catalysis

The use of turbulence modeling has further enhanced studies and reveals several pieces of information in catalysis. Several forms of research works have been conducted to contribute to the promotion of better understanding in the field of modeling catalytic reaction systems. Some of the recent works carried are reported, as shown below.

Christoph et al. [29] employed the computational fluid dynamics approach to undertake an empirical and numerical study of the air-hydrogen catalytically stabilized combustion (CST) over platinum (Pt) catalyst in a turbulent flow. This was to assess the viability of applying CST in different near-wall turbulence representative equations and gain insight into the combination of turbulence and hetero/homo-gaseous burning processes. The examination of the turbulence in the CST of the hydrogen-air mixture in the Pt-covered passages demonstrated the significance of near-wall turbulence equations. This captures the inducement of laminarization with strong-flow through the heat transferred via the catalyst filled walls. It was established that the size of turbulent transport was central to the appraisal of catalytic fuel conversion and the homogeneous ignition process.

The result obtained inspired further studies by Christoph et al. [29]. One of the studies was focused on forecasting with three separate low Reynolds number (LRe) turbulence representative equations (coiled directly from the literature). This finding showed that for specific turbulence representative equation, they have keen sensitivity for hetero-/homogeneous processes

Christoph et al. [29] further reported that a steady flow laminarization decreases the conversion of the fuel by use of catalyst, and it also aids the start of uniform ignition. So, a rise in the rates of turbulent transport gives rise to incomplete fuel combustion in the gaseous zone, leaked fuel conversion using catalysts, and finally, to flame extermination. However, the experimental studies of Christoph et al. [29] show that the model of Ezato reproduces with good accuracy both for the pre-ignition catalytic conversion and the onset of homogeneous ignition, although it overpredicted the degree of orderliness or laminarization mildly.

Rodrigo and Rosa [40] studied the use of CFD software for modeling turbulence in two or more phase flows inside TBRs under high pressure. However, the report shows that the use of CFD for single-phase reactors are a successful tool. So, the authors examined the use of CFD for exemplifying three-dimensional (3-D) interstitial flow in a reactor with two or more phases through the influence of fluctuating turbulent velocities and non-directional quantities for the multiphase (MTP) flow. In their studies, the effects of representative equations for turbulence in two or more phase flows were evaluated. An Euler-Euler representative equation was then evolved, and various RANs models such as realizable, standard, and RNG k-ε representative equations, including the RSM for the computer-aided simulation of hydrodynamics of the TBR under high pressure. The report showed that several analyses were executed for the study of the characteristics in the search for numerical solution parameters. This was because the

simulation accuracy frequently depended on the density of the mesh, sizes of the mesh, time interval, conditions for convergence, and schemes for unique identification. The various solutions for the hydrodynamic confirmation of the MTP flow model were compared, and it was identified that the CFD forecast with the monotonic upwind scheme for conservation laws (MUSCL) scheme agreed well with the data obtained from experiments. The findings were the results of the total variation diminishing (TVD) algorithm, which superseded the numerical dispersion displayed in the simulation of two or more-phase flows. The optimum conditions obtained were employed in the determination of the various RANS turbulence representative equations. Here, it was identified that a temperature rise resulted in evenness of fluid mal-distribution inside the packed bed. It confirmed that better agreement was achieved with standard k-ε and RSM representative equations for the multi-phase trickle bed reactors.

In 2010, Binxin [41] evaluated different turbulence models. These include the three (3) HRe k-ε representative equations, six (6) LRe k-ε representative equations, two (2) k-ω representative equations, and the RSM for a single-phase non-Newtonian fluid flow in anaerobic digesters. This was achieved by relating the pressure drops obtained from the study of the process's CFD. The findings from the simulation conducted unveil that the Chang-Hsieh-Chen account of the LRe k-ε representative equation gave superior efficiency compared to other representative equations in the prediction of the pressure losses. In contrast, the standard k-ω representative equation had a low computing cost and acceptable accuracy, which agreed with Rodrigo and Rosa [40].

Muhammad [30] simulated FCC riser using the model (k-e), three different mesh, and three lump kinetic models to studies the two-phase flow of vapor and catalyst to describe the profiles of temperature, mass fraction, and the gasoline product yield within the reactor (FCC riser). The finding shows that all the simulation results did not unveil any significant difference for different mesh sizes, which was not in agreement with the report of Rodrigo and Rosa [40] for the use of different mesh sizes to enhance accurate predictions.

It was further deduced that the wrong selection of kinetic models could also contribute to poor prediction. This was because Muhammad [30] deduced that the setback of three lump kinetic representative equations could not forecast the mole fraction of coke and low molecular weight gases separately. However, Muhammed [43] in 2015, further the study on the FCC riser, where the use of four lump kinetic models was employed to solve the problem reported in Mohammed [30]. Both Muhammad's studies [30, 43] predict that the most complex segment of an FCC rise is the inlet zone. Although, Muhammed [21] reported that the reaction holds in the initial 2–4

m of the length of the FCC riser, while Muhammed [43] reported that in the first 1 – 3 m. The difference in the prediction was due to the different kinetic models used. However, both works reported that the plant data and model results agree quite well, especially when more realistic kinetic parameters are incorporated [43, 44]. The model that is proposed finds usage in all FCC riser simulation processes.

Xiaomin et al. [39] modeled a 2D – CFD for a fixed-bed (FB) reactor that uses a catalyst to remove hydrogen from syngas. The study simulated the flow behavior across the bed of catalyst. The modeling involves the combination of the CFD model for a porous medium with a reaction kinetics model. This approach enables the study to identify a more efficient kinetics model, which could better predict the reaction profile in the presence of the catalyst. With the use of the reaction mechanism and a statistical test, the study was able to identify a reliable representative equation for kinetics with separation–adsorption of oxygen gas molecules, which indicates the rate-controlling step. This step was identified and validated. The CFD model of the porous medium, in addition to a representative equation for kinetics, could give results of a simulation that agrees well with data from empirical studies. It was observed that the space velocity had a lower influence on the reaction, with appreciable values of space velocity given in the range of 2000–2200 h^{-1}.

Also, the CFD representative equation was used to forecast the distributions of the primary reaction variable. These variables include the concentration of the reacting species and the temperature within the reactor. H_2 conversion and hotspot temperature were greatly influenced by the inlet gas temperature and less by the inlet H_2 concentration, which should be kept below 1.4 % to eliminate temperature runoff. The flow rate of O_2 in the reactor is a dependent variable that can ensure that the expected conversion and desirable hot spot temperature are achieved [39].

Dominick et al. [31] determined the extent of mal-distribution of flows within the bed of the catalyst using COMSOL Multi-physics and a 2-D axis-symmetric model. They also evaluated the influence of flow's direction, size of the catalyst pellet, total and different screen opposition, and the overall volume of flow passing inside the reactor. This study reveals that the catalyst with a diameter of 1/8" allowed increased bed velocity and flow rate at a specific pressure loss. However, a faster flow volume could be advantageous. The increased bed velocity implies that the residence time of the fluid in the catalyst bed will be reduced, and reaction conversion is minimized. Results also give values for the index of maldistribution that hints that the 1/16" catalyst gives superior distribution in the catalyst bed. Dominick et al. [31] indicated that the

particle size of the catalyst would be the critical factor in determining the distribution of flows in the catalyst bed center. An enhanced distribution within the catalyst bed positively promotes the conversion process and the reactor's efficiency. Hence, this revealed that the 1/16" catalyst is preferred in the modeled reactor since it promotes higher reactor efficiency.

Zhapbasbayev et al. [42] modeled the turbulent flow within a radial fixed bed reactor using the k-e model, employing a unified approach for writing the motion equation for the numerical solution of the problem in a hybrid region. However, the computational study signifies that the flow within the fixed bed results in the generation of the kinetic energy of turbulence fluctuation and its rate of gradual disappearance. The patterns of the kinetic energy of turbulence fluctuation and its rate of gradual disappearance show the similarity of all turbulent characteristics in terms of the Reynolds number estimation.

Afshin [32] evaluated the effect of carbon tetrachloride (CCl_4) addition as propulsion to the thermal cracking reactor due to the quantity of coke formed in various COT. The findings identified that the *k-ε* realizable model would best predict turbulence behavior in a cracking process. Arising from this research and in collaboration with other reports showed that the optimum volume of CCL_4 that ensures fixed conversion was 100 ppm, and beyond this value, the process will be uneconomical.

7. Conclusions

Turbulent flow is said to be a flow zone or space in fluid dynamics characterized by disorganized changes in velocity and pressure of the flow. It contrasts, a flow regime which holds when fluid flows in parallel layers, with negligible disturbance in the layers is referred to as laminar flow. Meanwhile, this study indicates that as the pore size of the catalyst becomes more extensive, the speed of reaction drops responsively. It can so be said that the larger the turbulence eddies in the reactant flow stream, the lower the speed of reaction. Also, this review shows that models like K–ε, and RSM turbulence model will be useful for the prediction or study of turbulence behavior in a catalytic reaction. It was identified that selecting the appropriate turbulence model in a kinetic study plays a vital role in promoting accurate reaction kinetics prediction when carrying out the simulation.

However, this study identifies that only a few research works give attention to the right or appropriate use or selection of a kinetic model for catalytic reaction systems. Besides, it was also identified that there are little works that employed the study of whether the S-a model could enhance better prediction in the modeling of turbulence in a catalytic reaction system.

8. Recommendations

It is recommended that further studies should consider the accuracy of predictions of catalytic reaction systems when using the *s-a* turbulence model. Attention may also be given to the study of the best approach to selecting the correct kinetic model of a catalytic system, and the effect of using incorrect or inappropriate kinetic models.

9. Acknowledgments

The first author wishes to acknowledge to support of the PTDF in making the program successful.

10. References

[1] G. K. Batchelor. *An introduction to fluid dynamics*, Cambridge, University Press, 2000.
[2] H. Tennekes, J.L. Lumley, L.J. Lumley. *A first course in turbulence*. Cambridge, MIT Press, 1996.
[3] F. C. K. Ting, J. T. Kirby. Dynamics of surf-zone turbulence in a spilling breaker, *Coastal Engineering*, vol. 27(3–4), pp. 131–160, 1996.
[4] E. Kunze, J. F. Dower, I. Beveridge, R. Dewey, K. P. Bartlett. Observations of biologically generated turbulence in a coastal inlet. *Science*, vol. 313 (5794), pp. 1768-1770, 2006.
[5] I. Eames, J. B. Flor. *New developments in understanding interfacial processes in turbulent flows.* Royal Society Publishing, 2018. http://rsta.royal societypublishing.org, accessed on 21 January 2018.
[6] S. H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Florida, CRC Press, 2018.
[7] H. G. Schuster, W. Just. *Deterministic chaos: an introduction*. Hoboken, John Wiley & Sons, 2006.
[8] S. T. Sie. Past, present, and future role of microporous catalysts in the petroleum industry. *Studies in Surface Science and Catalysis,* vol. 85, pp. 587-631, 1994.
[9] C. Bartholomew. Catalyst deactivation/ regeneration. *Encyclopedia of Catalysis,* 2002.
[10] C. H. Bartholomew, R. J. Farrauto. *Fundamentals of industrial catalytic processes*. John Wiley & Sons, 2011.
[11] Chemguide, Guide, https://www.Chemguide.co.uk/phys ical/basicrates/catalyst.html, accessed on 31 May 2018.
[12] T. Kuwabara, P. Kurzweil. Fuel Cells–Phosphoric Acid Fuel Cells | Anodes, *Encyclopedia of Electrochemical Power Sources,* p. 548-556, 2009.
[13] T. Oyegoke, F. N. Dabai, U. Adamu, B. E. Y. Jibril. Density functional theory calculation of propane cracking mechanism over chromium (III) oxide by cluster approach. *Journal of the Serbian Chemical Society,* vol. 85(5), pp. 1-14, 2020.
[14] H. Knözinger, K. Kochloefl. *Heterogeneous Catalysis and Solid Catalysts,* Ullmann's Encyclopedia of Industrial Chemistry, Wiley-VCH, Weinheim, 2002. DOI:10.1002/14356007.a05_313
[15] G. Vicente, M. Martınez, J. Aracil. Integrated biodiesel production: a comparison of different homogeneous catalysts systems. *Bioresource Technology,* vol. 92(3), pp. 297-305, 2004.
[16] T. C. Johnson, D. J. Morris, M. Wills. Hydrogen generation from formic acid and alcohols using homogeneous catalysts. *Chemical Society Reviews*, vol. 39(1), pp. 81-88, 2010.

[17] Z. Helwani, M. R. Othman, N. Aziz, J. Kim, W. J. N. Fernando (2009). Solid heterogeneous catalysts for transesterification of triglycerides with methanol: a review. *Applied Catalysis A: General*, vol. 363(1-2), pp. 1-10.

[18] T. Oyegoke, F. N. Dabai, U. Adamu, B. E. Y. Jibril. Insight from the study of acidity and reactivity of Cr_2O_3 catalyst in propane dehydrogenation: a computational approach. *Bayero Journal of Pure and Applied Sciences*, vol. 11(1), pp. 178-184, 2018.

[19] Y. Zhao, K. R. Yang, Z. Wang, X. Yan, S. Cao, Y. Ye, K. L. Materna. Stable iridium dinuclear heterogeneous catalysts supported on the metal-oxide substrate for solar water oxidation. *Proceedings of the National Academy of Sciences*, 201722137, 2018.

[20] S. Morales-Delarosa, J. M. Campos-Martin, J. L. Fierro. Chemical hydrolysis of cellulose into fermentable sugars through ionic liquids and antisolvent pre-treatments using heterogeneous catalysts. *Catalysis Today*, vol. 302, pp. 87-93, 2018.

[21] J. Gao, Y. Wang, Y. Du, L. Zhou, Y. He, L. Ma, Y. Jiang. Construction of biocatalytic colloidosome using lipase-containing dendritic mesoporous silica nanospheres for enhanced enzyme catalysis. *Chemical Engineering Journal*, vol. 317, pp. 175-186, 2017.

[22] D. C. Cubides-Roman, V. H. Pérez, H. F. de Castro, C. E. Orrego, O. H. Giraldo, E. G. Silveira, G. F. David. Ethyl esters (biodiesel) production by Pseudomonas fluorescens lipase immobilized on chitosan with magnetic properties in a bioreactor assisted by an electromagnetic field. *Fuel*, vol. 196, pp. 481-487, 2017.

[23] R. Das, M. Talat, O. N. Srivastava, A. M. Kayastha. Covalent immobilization of peanut β-amylase for producing industrial nano-biocatalysts: A comparative study of kinetics, stability, and reusability of the immobilized enzyme. *Food Chemistry*, vol. 245, pp. 488-499, 2018.

[24] J. Matthiesen, S. Wendt, J. O. Hansen, G. K. Madsen, E. Lira, P. Galliker, E. K. Vestergaard, R. Schaub, E. Laegsgaard, B. Hammer, F. Besenbacher. Observation of All the Intermediate Steps of a Chemical Reaction on an Oxide Surface by Scanning Tunnelling Microscopy. *ACS Nano.*, vol. 3(3), pp. 517–526, 2009.

[25] M. E. Davis, R. J. Davis. *Fundamentals of Chemical Reaction Engineering*, New York, McGraw-Hill Company, 2003.

[26] H. S. Fogler. *Essentials of chemical reaction engineering*. 4th Edition, Westford, MA, Pearson Education, 2006.

[27] B. Andersson. *Computational fluid dynamics for engineers*, Cambridge, Cambridge University Press, 2012.

[28] J. C. Ching, J. Sheng-Yuh. *Fundamentals of turbulence modeling*, Abingdon, Taylor & Francis, 1998.

[29] A. Christoph, J. Mantzaras, R. Schaeren, R. Bombach, A. Inauen, B. Kaeppeli, A. Stampanoni. An experimental and numerical investigation of homogeneous ignition in catalytically stabilized combustion of hydrogen/air mixtures over platinum. *Combustion and Flame*, vol. 128(4), pp. 340-368, 2002.

[30] A. Muhammad. Computational fluid dynamics (CFD) prediction of mass fraction profiles of gas oil and gasoline in fluid catalytic cracking (FCC) riser. *Ain Shams Engineering Journal*, vol. 3(4), pp. 403-409, 2012.

[31] T. I. Dominick, I. E. Specht. *Experimental and Numerical Analysis of Flow Mixing in Packed Beds*, Doctoral Thesis, Otto von Guericke University, Magdeburg, 2016.

[32] D. Afshin. Evaluation of Adding Carbon Tetrachloride as Propulsion to the Thermal Cracking Reactor due to the Amount of Formed Coke in Different Coil Outlet Temperatures (COT). *J. Thermodyn. Catal.*, vol. 8(179), pp. 2, 2017.

[33] W. Rodi. *Turbulence models and their application in hydraulics*. Florida, CRC Press, 2017.

[34] J. Sodja. *Turbulence models in CFD*, Technical Report, University of Ljubljana, Ljubljana, pp.1-18, 2007.
[35] S. S. Girimaji. Pressure–strain correlation modeling of complex turbulent flows. *Journal of Fluid Mechanics*, vol. 422, pp. 91-123, 2000.
[36] C. Stephen, D. Caraeni, L. Fuchs. Large-eddy simulation of the flow through the blades of a swirl generator. *International journal of heat and fluid flow*, vol. 21(5), pp. 664-673, 2000.
[37] J. Lumley, G. Newman. The return to isotropy of homogeneous turbulence, *Journal of Fluid Mechanics*, vol. 82, pp. 161–178, 1977.
[38] A. Mishra, S. Girimaji. Intercomponent energy transfer in incompressible homogeneous turbulence: multi-point physics and amenability to one-point closures, *Journal of Fluid Mechanics*, vol. 731, pp. 639–681, 2013.
[39] X. Xiaomin, J. Dai, Z. Luo. CFD modeling using heterogeneous reaction kinetics for catalytic dehydrogenation syngas reactions in a fixed-bed reactor. *Particuology*, vol. 11(6), pp. 703-714, 2013.
[40] L. J. Rodrigo, F. M. Rosa-Quinta. Turbulence modeling of multiphase flow in high-pressure trickle-bed reactors. *Chemical Engineering Science*, vol. 64(8), pp. 1806-1819, 2009.
[41] W. Binxin. Computational fluid dynamics investigation of turbulence models for non-Newtonian fluid flow in anaerobic digesters. *Environmental science & technology*, vol. 44(23), pp. 8989-8995, 2010.
[42] U. K. Zhapbasbayev, G. I. Ramazanova, O. B. Kenzhaliev. Modeling of turbulent flow in a radial reactor with a fixed bed, *Thermophysics and Aeromechanics*, vol. 22(2), pp. 229-243, 2015.
[43] A. Muhammed. Prediction of gasoline yield in a fluid catalytic cracking (FCC) riser using k-epsilon turbulence and 4-lump kinetic models: A computational fluid dynamics (CFD) approach. *Journal of King Saud University-Engineering Sciences*, vol, 27(2), 130-136, 2015.
[44] A. Gianetto, H. I. Farag, A. P. Blasetti, H. I. de Lasa. Fluid catalytic cracking catalyst for reformulated gasoline. Kinetic modeling. *Industrial & engineering chemistry research*, vol. 33(12), pp. 3053-3062, 1994.

Environmental Sustainability Study of Refinery Wastewater System Using Pinch Technology

Lawrence C. Edomwonyi-Otu,
Department of Chemical Engineering, Ahmadu Bello University Zaria, Nigeria
Department of Chemical and Petroleum Engineering, Delta State University, Nigeria
uceclce@ucl.ac.uk, edomwonyi-otu@delsu.edu.ng
and
Isaac Okewu,
Department of Chemical Engineering, Ahmadu Bello University Zaria, Nigeria
isaacokewu@yahoo.com
Paul C. Okonkwo,
Department of Chemical Engineering, Ahmadu Bello University Zaria, Nigeria
chemsprom@gmail.com

Abstract: Pinch technique was employed to study the potential of environmental sustainability in a Nigerian Refinery wastewater system. This was achieved by considering nine major water using units: cooling tower, ion exchangers, laboratory, SWS Desalter, FCC SWS, caustic treating, in-house water, fire service and boiler blowdown. In each of the operations, the water flowrate, and the inlet and outlet contaminants concentration were determined. The Pinch analysis software plotted the concentration interval diagram for quick determination of the mass load, cumulative mass load and the pinch point. The water pinched at BOD of 110ppm with average outlet contaminant concentration of 210ppm. About 50% reduction was observed for required fresh water flowrate after the water reuse optimization. Wastewater generation was also reduced by 38% considering reuse technique and 49% for regeneration reuse option. The final regeneration outlet concentration obtained was found to be 48ppm and within acceptable environmental standard.

Keywords: Pinch technology, wastewater, sustainability, optimization, refinery, pollution

دراسة الاستدامة البيئية لنظام الصرف الصحي في المصفاة باستخدام تقنية القرص

الملخص: تم استخدام تقنية القرص لدراسة إمكانات الاستدامة البيئية في نظام الصرف الصحي النيجيري لمصفاة التكرير. وقد تم تحقيق ذلك من خلال دراسة تسع وحدات رئيسية تستخدم المياه: برج التبريد، والمبادلات الأيونية، والمختبر، ومحلل SWS، وFCC SWS، ومعالجة المواد الكاوية، والمياه الداخلية، وخدمة الحريق، وتطهير سخانات المياه. في كل عملية من العمليات، تم تحديد معدل تدفق الماء وتركيز ملوثات المدخل والمخرج. قام برنامج تحليل القرص برسم مخطط فاصل التركيز لتحديد سريع للحمل الكتلي وحمل الكتلة التراكمي ونقطة الضغط. الماء المضغوط عند الطلب البيوكيميائي للأوكسجين ١١٠ جزء في المليون مع متوسط تركيز ملوث للمخرج ٢١٠ جزء في المليون. لوحظ انخفاض بنسبة ٥٠٪ تقريبًا في معدل تدفق المياه العذبة المطلوبة بعد تحسين إعادة استخدام المياه. كما تم تقليل توليد المياه العادمة بنسبة ٣٨٪ بالنظر إلى تقنية إعادة الاستخدام و٤٩٪ لخيار إعادة استخدام التجديد. وجد أن التركيز النهائي لمخرج التجديد الذي تم الحصول عليه هو ٤٨ جزء في المليون وضمن المعايير البيئية المقبولة.

1.Introduction

Water Pinch analysis can be applied for the optimal and more beneficial usage of process water (including wastewater) in process utility. It is the mass analogy to optimization of heat exchangers network via thermal Pinch analysis (Linnhoff et al., 1990). The idea of water pinch technology has been bought by many industries; especially oil refineries and the report have it that the volume of fresh water demand has been reduced from 60% to approximately 25% (Abubakar, 2016). The graphical the rigorous mathematical optimization approaches are usually applied in Pinch analysis processes. Pinch analysis consists mainly of two steps which are targeting and design. Targeting is a method of locating the water pinch, the minimum fresh water flow rate, and this is done before the network design or modifications. Composite curves/source and sink/purity profiles as well as excess water diagram with water cascade as a supplement can be displayed in a graph with the help of Pinch software (Wang and Smith, 1994).

Fresh water usage and capacities alongside the resultant wastewater in industries is being minimized with the help of Pinch analysis. This is important because it helps companies through systematic and technical analysis of water to meet the ever-increasing demand for fresh water and the stringent regulations for treatment and release of wastewater to the environment. The optimal reuse and regeneration of water including opportunities for effluent treatment can be identified with advanced algorithms in Pinch analysis. Reduction in feedstocks as well as recovery of economic products in outlet streams have been attributed to use of Pinch analysis (Lahnsteiner et al., 2012).

According to (Wang & Smith, 1994), wastewater optimization can be carried out in four ways;

1. **Changes in Process:** This involve process reconfiguration aim at inherent water demand minimization, a clear illustration of this is the conversion of wet cooling tower to dry air coolers.
2. **Water Reuse:** Water effluent from one process can be used in other process without any treatment if the quality of that water is considered fit for the other operations without compromising the quality requirement of the operations. This technique dose not only reduced the volume of the fresh water require but also reduce the volume of the waste water generated which in turn reduce the environmental contaminations and also reduction in overall operational cost.
3. **Regeneration reuse:** Wastewater can be regenerated by subjecting it to one or two water treatment processes to completely or partially reduce the level of contamination that

made it unfit to be reused in other water using processes. The mass-load of contaminants as well as the volumes of required fresh water and wastewater can be minimized by regeneration. (Khezri et al., 2010)

4. Regeneration Recycle: Wastewater quality can be improved by regeneration by suggesting the waste water through one on two treatment processes so that it can be recycled. In regeneration recycle, the regenerated water may be reused in other water using activities where the water stream was previously used. The introduction of the concentration interval diagram as well as the concentration composite curve, the concentration interval diagram and the freshwater pinch usually helps with the analyses. The work of Wang and Smith (1994) was used to calculate the lowest flow rates of both fresh inlet water and that after regeneration process.

In a typical oil refinery, over 2,000 products can be extracted from crude oil by using several unit operations including distillation, cracking, alkylation, in addition to polymerization, coking, hydrotreating, and others. In most of these processes, wastewater is been generated daily in the refining process which is based on the operations that produce the polluted water. Wastewater is mainly grouped as process and non-process types. The wastewater from the refining operations that has been in touch with the hydrocarbons is called is known as the process type while those from surface water runoffs, washing of machines, storage and office facilities as well as cooling towers is called non-process wastewater (Al-Shamrani et al., 2002). The streams usually given separate treatment for best outcome for the avoidance of pollution by harsh contaminants in substantial volume of water (Peng et al., 2008). Process wastewater is the result of the presence of water as suspension or emulsion when crude oil is produced from oil wells (particularly aging ones) and transported to the refinery (Edomwonyi-Otu and Angeli 2019, 2015). Moreover, liquid or gaseous water is also used specifically for enhanced oil recovery processes in aging wells (Cheryan and Rajagopalan, 1998). The application of gaseous and liquid water as a diluent and as a stripping fluid in distillation, cracking and desalting operations contributes to the amount of process wastewater in refineries (IPIECA, 2010).

The formation of oil layer in water bodies resulting from discharged wastewater that contains oil and grease is a source of danger to the aquatic life. This oil layer reduces the amount of light that can penetrate to maintain production of oxygen and other photosynthetic activities. The population of the ecosystem is also negatively impacted as the amount of oxygen that can be

dissolved from the surrounding air is drastically reduced by the oil layer. Nacheva (2011) and Chao & Liang (2008) further confirmed that the health of the environment and life in general is at risk because of the poisonous nature of the compounds in the oil-contaminated water that has been unabatedly discharged to the environment. Some of the compounds in the oil-contaminated water include phenols, benzene, ethylbenzene, xylenes, sulfides, as well as suspended solids (SS), ammonia and other polyaromatic hydrocarbons (PAH)(IPIECA, 2010). The chronically toxic nature of phenols when consumed even in small amounts has bee known to result in impairment of sight, diarrhea, mouth souring and dark urine people in addition to life threatening consequences for aquatic life (Andreozzi et al., 1999). More worrying is the fact that the solubility of the compounds in water can serve as a ready media for the synthesis of even more toxic substances (Guerra et al., 2011). Bathing or showering with oil-contaminated water can result in the absorption of benzene, toluene, ethylbenzene and xylenes (BTEX) into the body through the skin leading to anemia and cancer. The presence of ethylbenzene, xylene and toluene in the body can cause improper functioning and damages to the liver, kidney and the nervous system (Peng et al., 2008).

The usual refinery practice for wastewater discharge to water bodies follows a two-step process to eliminate suspended oil from the accumulated wastewater. This is accomplished with the help of API separators and then the dissolve air flotation (DAF) unit. The equalization system receives the water from the secondary oil/water separation unit and levels out its concentration. The effluent from the equalization unit is then fed to the aeration tank clarifier for biological treatment and finally to tertiary treatment before release (Guerra et al., 2011). Products from the refinery include; liquefied petroleum gas (LPG), automotive gas oil (AGO), premium motor spirit (PMS), kerosene, fuel oil, sulfur. The lubrication oil complex produces base oils, asphalt/bitumen and waxes (KRPC, 2002) as additional products.

1.1 Objectives

The objective of this work is the application of water pinch analysis to Kaduna Refining & Petrochemical Company (KRPC) water management system. This system will reduce the KRPC freshwater demand that has resulted in a lot of water diversion from the Kaduna River. It will also reduce the amount of wastewater generation which are usually discharged into the environment without due regard to the attendant consequences.

2. Materials and Methods

2.1 KRPC Water Supply System: The following steps were taken to carry out the pinch analysis: KRPC water distribution network was analyzed from their water facility beside River Kaduna at Kamazo. The water then undergoes screening to remove solids debris, wooden piece etc, and thereafter sent through 13km network of pipes to a 750m^3 reservoir. It is then sent to clarifiers A, B and C, and then to the high rate filters, cooling tower, service water, ion exchange and to the activated filtration system before it is certified to be drinkable. The flowrate of water to the cooling tower, ion exchange, laboratory, desalter, fluid catalytic cracker (FCC), caustic treating, in-house, fire service and boiler blow down were obtained. The inlet and outlet water samples were collected from each of the unit using cleaned fresh plastic bowls that were properly rinsed with both tap and distilled water at least twice. The samples were then labeled appropriately and transported to the laboratory where it was refrigerated at 4^0C prior to analysis.

The parameters of the raw water determined were chemical oxygen demand (COD), conductivity, total hardness, biological oxygen demand (BOD), turbidity and total solid (TS) (Edomwonyi-Otu and Adelakun 2018). For the purpose of this study, BOD was selected as the limiting parameter. BOD is the measure of oxygen required to oxidize the organic matter present in sample water. The result obtained for BOD was used in the water pinch software developed by James Manna of Virginia Polytechnic Institute and State University. Pinch point, mass load, cumulative mass load, concentration interval diagram, concentration composite curve, preliminary network design and evolution network design were determined and constructed with the aid of the water pinch software. The experiments were carried out at a Civil Engineering Water Laboratory to improve the quality of treated effluent and to determine the appropriate treatment scheme for recycle/reuse of treated water. The treatment routes investigated were sand filtration and activated carbon column. The experimental setup for sand filtration and activated carbon column process was assembled in the same laboratory to carry out the recycle studies on effluents.

2.2 Regeneration and Recycling Process

2.2.1 Sand filtration: The sand used as a filter media was screened between 0.8 - 1.2mm particle size. The sand was thoroughly washed with tap water to remove all the dirt: both organic and inorganic matter, until the water became clear. The washed sand was oven-dried (Gallenkamp TM OV-420) at 100°C to remove the moisture/water content. The filtration experiment was carried out on secondary treated effluent of KRPC wastewater in a bench scale laboratory model (batch mode). The sand filter consisted of a circular glass tube with a packed bed of sand and gravel. The size of column was 8cm diameter by 78cm height. The reason for the choice of the size of the sand used was to ensure that rate of filtration was not too slow. The size of the sand at different layers is as seen in Table 1. The sand was placed on wool and perforated plate. The sand filter was fed from top through gravity with secondary treated effluent and the filtrate was collected at the bottom. The filtered effluent was analyzed for pH, SS, COD, BOD, chloride, nitrate, alkalinity, turbidity, total dissolved solids (TDS), TS, oil and grease and some heavy metals.

Table 1: Sand filtration

Layer	Size (mm)	Height (cm)
Top	0.8	6
Middle	1.2	6
Bottom	2.0	6

2.2.2 Granular activated carbon (GAC): The granular activated carbon used for this work was obtained from Alpha Chemicals in Kaduna State, Nigeria. The activated carbon was washed with tap water to remove impurities. The washed activated carbon was oven-dried (Gallenkamp TM OV-420) at 100°C to remove the moisture/water content. The adsorption test was conducted on sand filtrate in a bench scale laboratory model (batch mode). The filtrate effluent was also analyzed as in Section 2.2.1.

2.2.2.1 Adsorption capacity of GAC: The batch adsorption experiment was performed by contacting 1.6g, 2.8g, 4.0g, 4.8g, and 6g of GAC in 160ml of filtered effluent in 5 sets. Finetech Magnetic stirrer at 150rpm was used to provide continuous stirring of the mixture for the study duration for improved mixing and to enhance the degree of contact. Each set was removed from the shaker after 40 minutes and pH of 7 to 8 (Cheng 2008; Zheng et al., 2018). The completely mixed activated carbon-effluent system was passed through the Whatman filter paper. The filtrate effluent was also analyzed as in Section 2.2.1 to determine optimum dose.

2.2.2.2 Determining the effect of contact time on adsorption capacity: Samples of 160ml wastewater were added to the five different GAC dose (4g) in 250 ml Pyrex flasks. The contents were then agitated at a constant speed for different contact times of 12, 24, 36, 48, and 60 minutes. The filtrates were collected separately and also analyzed as in Section 2.2.1 to determine optimum contact time.

3. Results and Discussions

3.1 Assessment of unit flow rate and quality requirement : To carry out the pinch analysis, BOD was analyzed across the nine major water unit operations. BOD was considered because it is one of the most important water quality parameters, and it helps in the determination of the amount of organic matter in a given sample of water. It also provides useful clue on the effect discharged wastewater will have on the receiving environment. Table 2 shows the change in BOD across the nine operations considered. It can be observed that operations such as desalters, FCC, and caustic treating units generated effluent with higher content of BOD, while operations such as boiler blowdown, ion exchanger, laboratory and in-house water, contributed far less to BOD in the effluent water. The limiting inlet value of BOD was compared to the design specification and as can be seen when BOD as a single contaminant is considered, KRPC wastewater effluent agrees with the design specification (KRPC, 2002).

Table 2 Limiting water data for BOD

Operations	Limiting flowrate (te/hr)	Limiting inlet concentration (ppm)	Limiting outlet concentration (ppm)	Mass load (kg/hr)	Design specification
Cooling water (Operation 1)	96	45.00	94.00	4.90	100
Ion Exchangers (Operation 2)	178	17.00	110.00	16.64	50
Laboratory (Operation 3)	12	13.00	74.00	0.38	50
SWS Desalter (Operation 4)	29	94.00	420.00	9.48	180
FCCSWS (Operation 5)	10	94.00	380.00	2.86	180
Caustic treating (Operation3) (Operation 6)	10	94.00	410.00	2.40	180
In house water (Operation 7)	7	7.00	18.00	0.42	95
Fire service (Operation 8)	38.10	86.00	420	3.40	180
Boiler blow down (Opera 9)	50	4.0	9.0	0,18	0

Fig. 1 is the concentration interval diagram (CID) for the limiting data for BOD in Table 2. The CID was constructed by matching the inlet and outlet contaminant concentrations with the flowrate of water across the nine operations. CID allowed for easy determination of mass load and subsequently the cumulative mass load. From the CID constructed, the flow rate of the fresh water requirement was reduced from 430 te/hr to 214 te/hr which was in agreement with the similar work carried out on Tehran Oil Refinery by Mohammadnejad et al., (2011). The concentration composite curve (CCV) of Fig. 2 was constructed by using the cumulative mass load generated from concentration interval diagram of Fig. 1. The pure water line was plotted with an assumption that inlet concentration of the water is equal to zero and the outlet concentration is maximal. The pinch point indicated at the point where the pure water line made a tangent to the concentration curve. This is indicated at 110ppm which was in agreement with the KRPC design specification. The pinch point is also the concentration critical point, and above this value, water is discouraged from been used (Ipsital 2010).

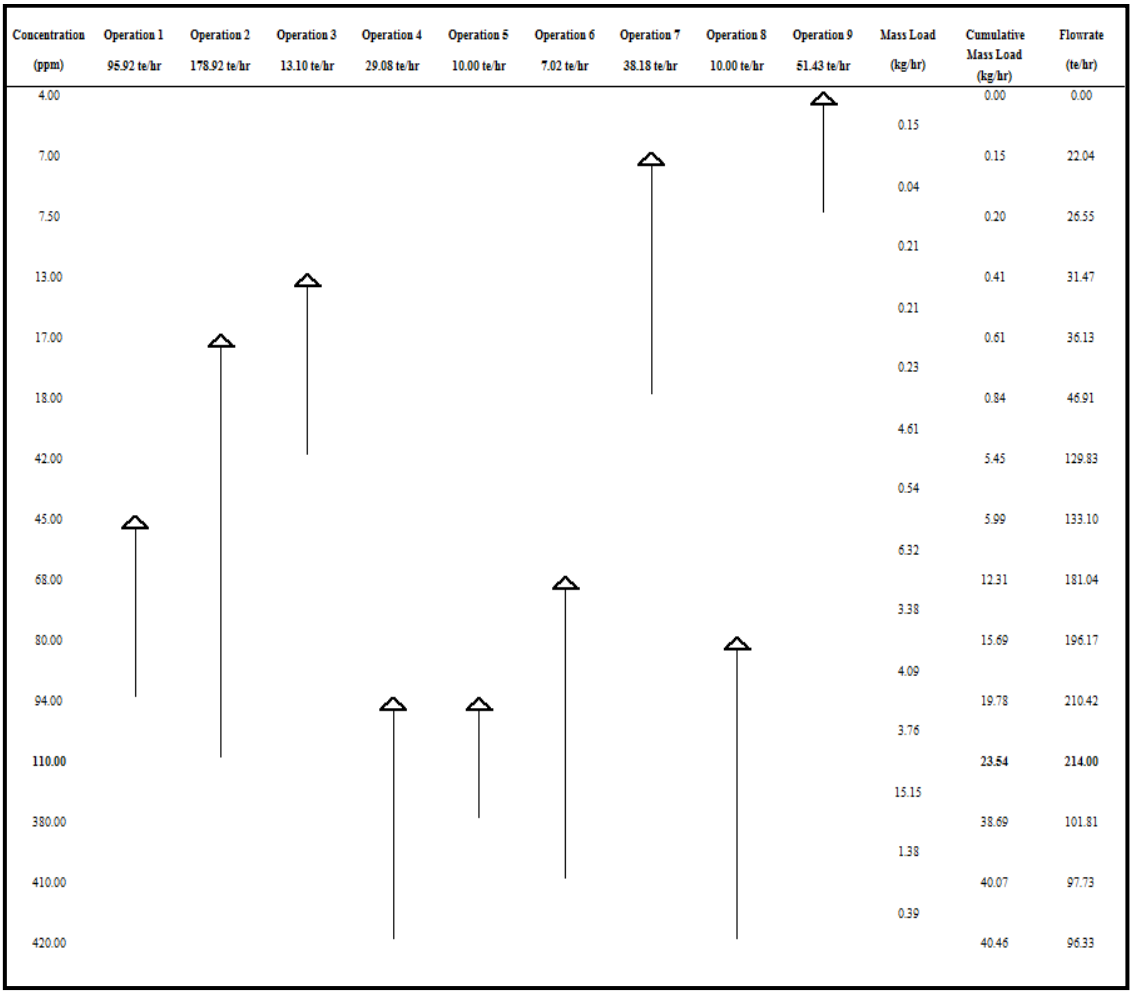


Figure 1. Concentration Interval Diagram for BOD

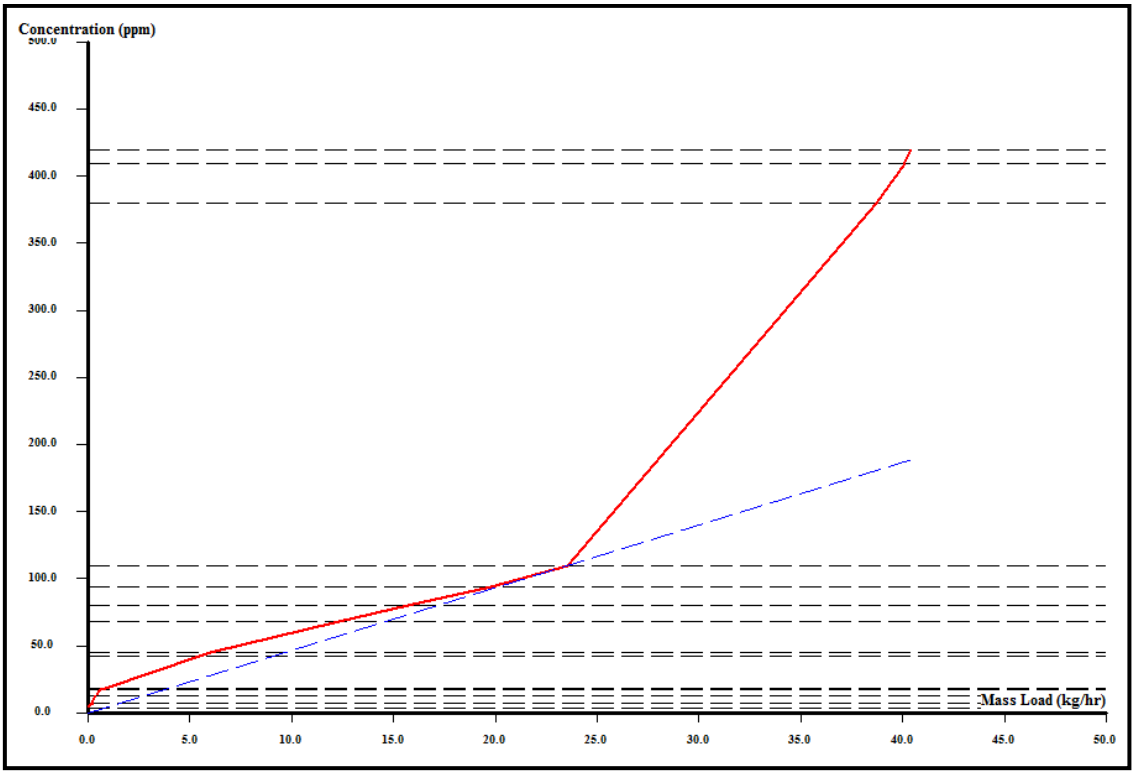


Figure 2. Concentration Composite Curve for BOD

Fig.3 is the evolution networks design for limiting data of Table 1. It was constructed from the concentration interval diagram of Fig. 1 and the concentration composite curve of Fig.2. It shows that operations one to nine makes use of the fresh water but the sources from the operations 2, 3, 7 & 9 formed the major sink to operations 1, 4, 5, 6 & 8. Hence, this provides an explanation for the 50% reduction in the volume of the fresh water need. Fig.3 also showed final networks for BOD as a single parameter. It can be seen that operation 3 required 50te/hr of fresh water before optimization. However, after optimization the volume of fresh water required was completely replaced by the effluent from operations 2, 3 & 9.This is due to the fact that the outlet of operations 2, 3 & 9 are less contaminated and the average outlet concentration was below the pinch point and can therefore be reused (Bush and Frankz, 1996). Operation 2 also had 30% of its fresh water requirement replaced by the outlet of operation 1, 7 & 9. Operation 3 requires 9.05te/hr of fresh water as its only source since no other outlet from any operation is reusable when BOD is the only parameter considered. Operations 4, 5, 6, & 8 have 50% of their required fresh water replaced by flows from operations 1, 2, 3, 7 & 9, and their outlet is highly contaminated and hence are not fit for reuse in any other operation (Mohammadnejad et al., 2011).

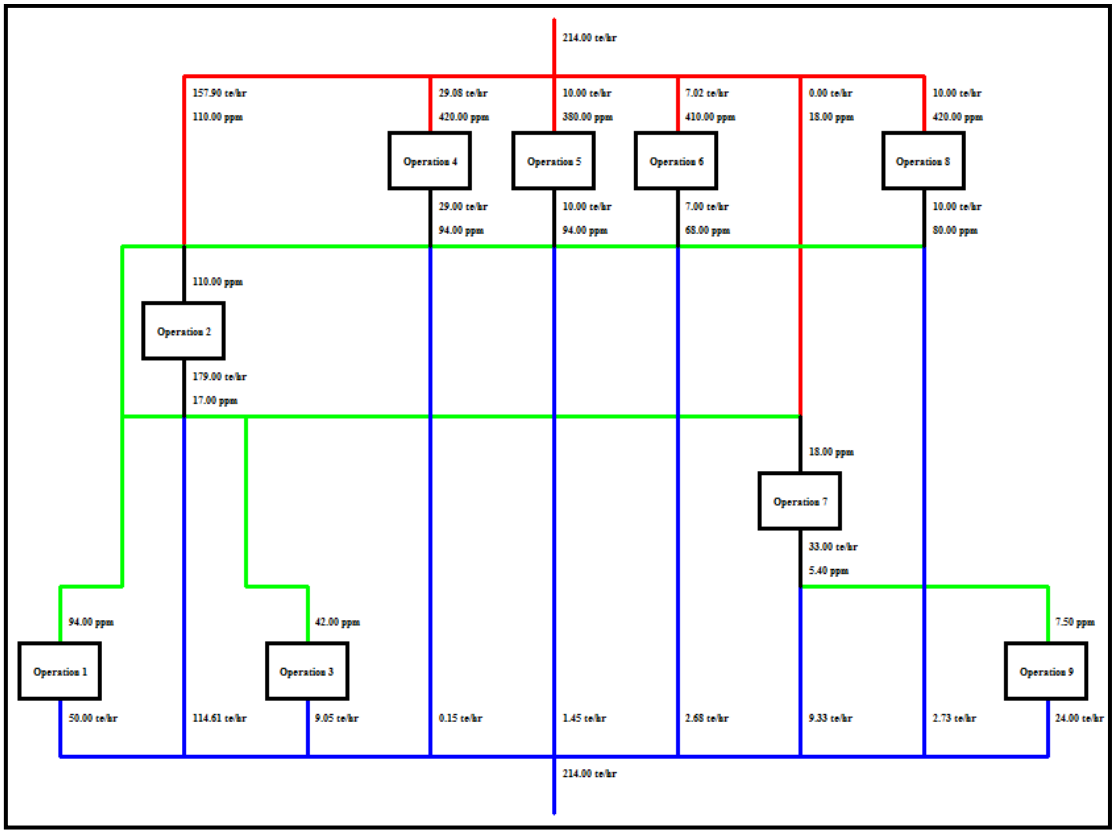


Figure 3. Network Evolutions for BOD

To determine the effect of regeneration on the environment, the wastewater was subjected to sand filtration process after which the filtrate was treated with activated carbon. Table 3 shows the physicochemical parameters and quality of the treated effluent water leaving/discharged from KRPC to Romi River, Kaduna.

Table 3 Physicochemical parameter of the effluent

S/No	Parameters	Value
1	Temperature (T^0C)	27.70
2	pH	7.60
3	Turbidity (NTU)	63.20
4	Total solid (mg/l)	450.10
5	BOD_5 (mg/l)	430.00
6	COD (mg/l)	720.65
7	Oil and grease (mg/l)	15.02
8	Total hardness (mg/l)	71.02
9	Conductivity (µs/cm)	400.63

It can be seen that the effluent from KRPC did not meet up with the discharge guideline by National Environmental Standard and Regulation Enforcement Agency (NESREA) as shown by turbidity, oil and grease content of the effluent water. Hence, the need to investigate the suitable processes to make the wastewater meet the NESREA guidelines for effluent discharge (NESREA 2010).

3.2 Sand filtration: The effluent water parameters were compared with the design specification and it was observed that the effluent water can only be used for fire services which requires water of lower quality. However, this is rarely used because fire outbreaks are not common hence the need to subject the effluent to further treatment so as to enhance it usability.

Table 4 Physicochemical parameter of the Sand filter effluent

S/No	Parameters	Experimental Value	Design specification	NESREA
1	Temperature(T^0C)	26.50	30	
2	pH	7.80	8	6 - 9
3	Turbidity(NTU)	9.90	20	< 40
4	Total solid(mg/l)	147.50	200	2000
5	BOD_5(mg/l)	205.50	50	30
6	COD(mg/l)	573.00	100	1000
7	Oil and grease(mg/l)	5.60	10	10
8	Total hardness (mg/l)	58.12	180	-
9	Conductivity(µs/cm)	260.00	110	400

Table 4 shows the performance of the sand filter which consists of a circular glass tube that was filled with sand of different sizes. The glass tube used was 5cm wide and 78cm high, and was filled with sand of 2mm, 1mm and o.8mm diameter respectively to a height of 6mm. The filter is categorized as slow filter since the flowrate is less than 0.017gpm (Cheryan and Rajagopalan, 1998). It was observed that the filter media reduced the turbidity from 63.20 to 9.90 NTU, equivalent to 84% turbidity removal while the TS and grease contaminant concentration were reduced by 72 and 80% respectively. Similarly, the COD and BOD removal were 20 and 19% respectively, as shown in Table 4. The sand filter effluent was compared with the KRPC water design specification and it was observed that the water was now fit for reuse for other service purposes including fire service. The BOD content was too high to be fit for use in most of the operations and hence, the sand filter effluent was subjected to activated carbon treatment.

3.3 Activated Carbon: The result of the activated carbon treatment is shown in Table 5. As can be seen, the COD, conductivity, total hardness, BOD, turbidity and total solid contaminant concentration were reduced by 64, 39, 65, 52, 28, and 39% respectively. The result obtained was within acceptable limits and corroborates the report by Amuda and Ibrahim, (2006), who treated industrial wastewater by using natural adsorbents.

Table 5: Adsorption capacity of GAC

S/No	Parameters	1.6g (GAC)	2.8g (GAC)	4g (GAC)	4.8g (GAC)	6g (GAC)
1	Temperature(T^0C)	27.50	28.00	28.10	28.00	28.02
2	pH	7.60	7.50	7.20	7.50	7.50
3	Turbidity(NTU)	7.48	7.10	5.10	5.10	5.98
4	Total solid(mg/l)	105.04	103.00	75.10	74.25	75.00
5	BOD_5(mg/l)	101.00	93.00	54.45	44.03	43.06
6	COD(mg/l)	216.04	210.00	80.0.65	80.5.22	89.94
7	Oil and grease(mg/l)	1.79	1.47	1.02	1.00	1.00
8	Total hardness (mg/l)	23.32	20.99	18.59	18.60	18.60
9	Conductivity(µs/cm)	194.00	184.00	172.00	174.00	173.04

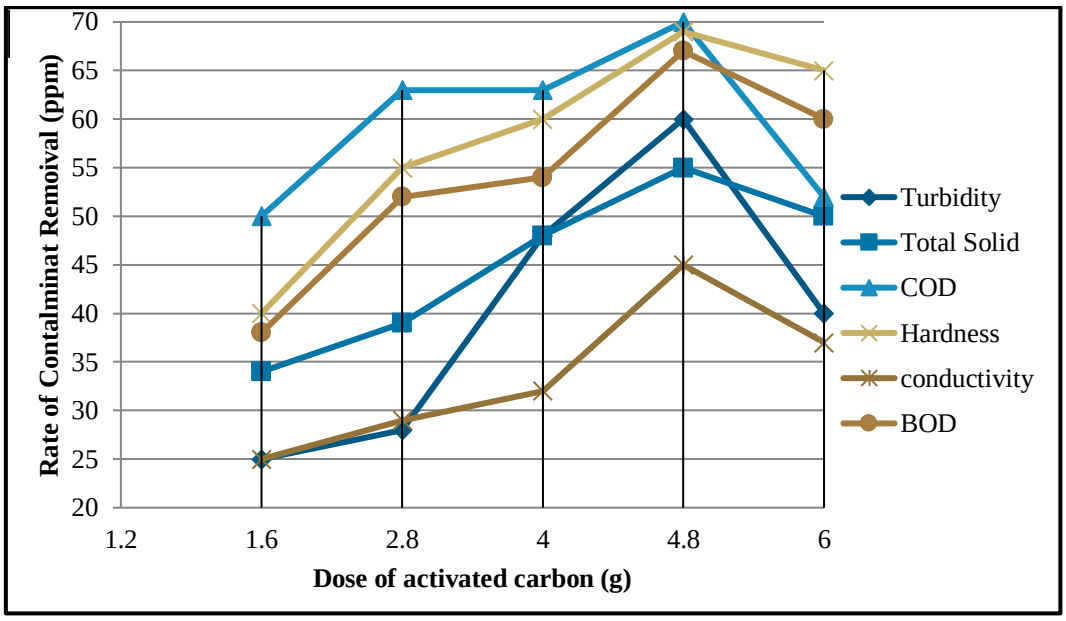


Figure 4. Adsorption capacity of GAC

Fig.4 presents the removal of COD, conductivity, total hardness, BOD, turbidity and TS. The dose was varied for 60 minutes (Couper et al., 2010). It can be seen that the percentage contaminant removal increased with the dose (g), particularly from 4g, to an optimum of 4.8g. This may be attributable to an increased amount of the replaceable sites and increased surface area (Edison et al., 2011). However, as the dose increased beyond 5g, the absorption rate started to decline. An evidence of saturation of the pores of the activated carbon.

3.3.1 Contact time: To obtain the optimum contact time of the activated carbon on the wastewater, batch experiment was conducted at different times; ranging from 12 minutes to 60 minutes (Bessarabov and Twardowski, 2002). Fig. 5 shows contact time versus absorption capacity of activated carbon. COD and BOD are our main targets for the activated carbon treatment since effluent from sand filter would have been reusable in most of the processes except that their concentration was too high. As can be observed, the optimum contact time was 36 minutes for all parameters and particularly the COD and BOD. While the COD removal decreased sharply after 24 minutes, the BOD removal decreased steadily and both parameters plateaued from 36 minutes. This may be due to the saturation of the pores of the GAC and it thereafter attained equilibrium (Edomwonyi-Otu et al., 2013; Cheng 2008).

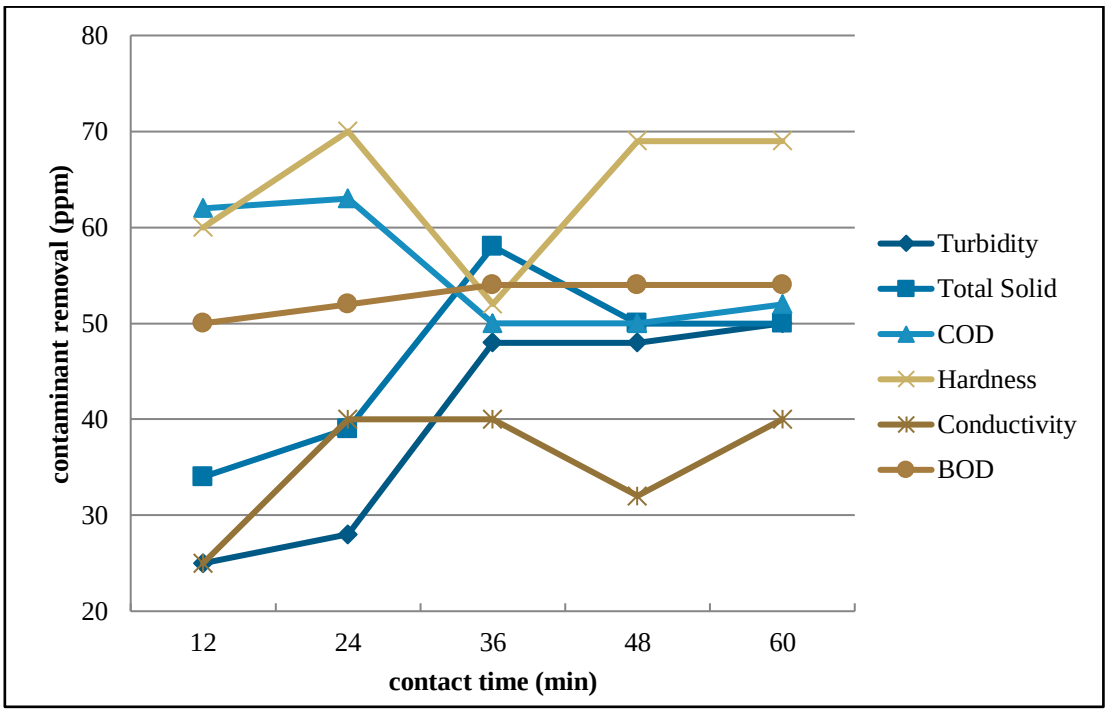


Figure 5. Contact time of wastewater with activated carbon

In pinch analysis, the recommended regeneration concentration C_0, is usually taken to be the pinch point (Lahnsteiner et al., 2012). In this work, two regeneration processes were carried out; sand filtration and activated carbon. The combination of these processes can increase savings by further reducing the volume of fresh water demand by 15%. From the regeneration processes carried out, the optimum regeneration concentration for the limiting parameter, BOD was 50ppm. This was added to initial flowrate for each unit operation and matched with each contaminant concentration to construct CID in Fig.6. This was used to determine the flowrate at 162.95te/hr and pinch point at 110ppm.

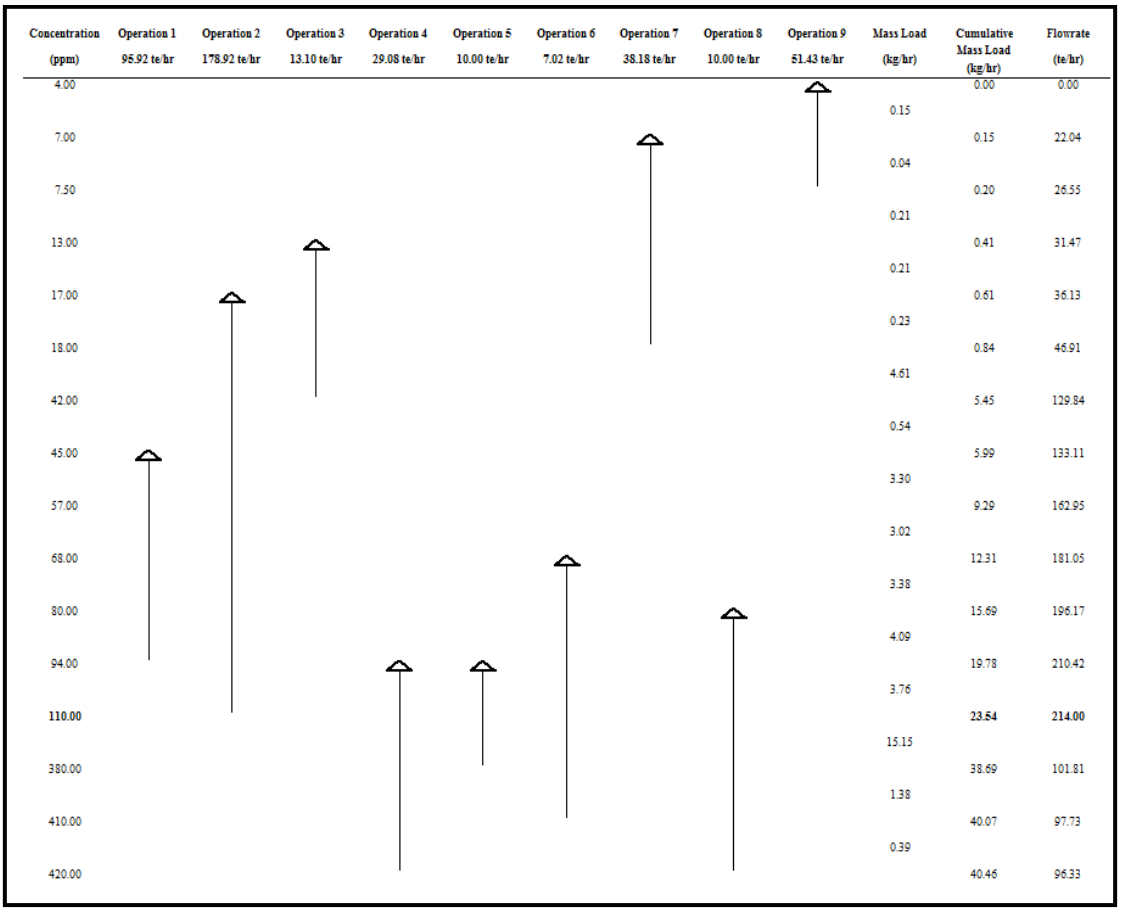


Figure 6. CID for BOD Regeneration

Fmin = 162.95te/hr; Freshwater pinch at 110.00 ppm; Average outlet concentration is 254.58 ppm

4. Conclusions

The application of water pinch analysis as a tool for environmental sustainability was demonstrated on Kaduna Refining and Petrochemical Company (KRPC) water system. Water reuse and regeneration approaches were considered. The water reuse approach showed

reduction of fresh water demand from 430te/hr to 214.86te/hr equivalent to 50% reduction in the volume of fresh water required with an average outlet concentration of 201ppm. After consideration of the regeneration technique, the wastewater was reduced by 38%, and 49% for regeneration reuse option. This resulted in reduced cost of wastewater treatment and environmental contamination. The final regeneration outlet concentration was found to be 48ppm and falls within the limit of environmental regulation as designated by NESREA (40 - 50ppm, NESREA 2010). Hence, the water reuse and regeneration approaches are recommended for KRPC water supply system for the purpose of optimal fresh water utilization and environmental sustainability.

References

[1] Abubakar, S. (2016). Optimization of Water Reuse Network Using Water Pinch Technology (WPT) by Considering Single Contaminant for Kaduna Refinery. *African Journal of Biotechnology*. 5(16), 1483-1487

[2] Al-Shamrani, A.A., James, A., and Xiao, H. (2002). Destabilization of Oil-Water Emulsions and Separation by Dissolved Air Flotation. *Water Research,* 36(6), 1503-1512.

[3] Amuda O.S., and Ibrahim A.O. (2006). Industrial Wastewater Treatment Using Natural Material as Adsorbent. *African Journal of Biotechnology*. 5(16), 1483-1487.

[4] Andreozzi, R., Caprio, V., Insula, A., and Marotta, R. (1999). Advanced Oxidation Processes (AOP) for Water Purification and Recovery. 53(1), 51-59.

[5] Bessarabov, D., and Twardowski, Z. (2002). Industrial Application of Nanofiltration - New Perspectives. *Membrane Technology*, 6-9. *Journal of Environmental Engineering*. 128(11), 1018-1029.

[6] Bush, K.E., and Frantz, T. (1976). Refinery Wastewater Treatment and Reuse. 13–15. In citeseerx.ist.psu.edu.

[7] Chao, Y., and Liang T.M.A. (2008). Feasibility Study of Industrial Wastewater Recovery Using Electrodialysis Reversal. *Desalination.* 221(1-3), 433-439.

[8] Cheng, W. (2008). Adsorption Characteristics of Granular Activated Carbon and SPME Indication of VOCs Breakthrough. *Aerosol and Air Quality Research.* 8(2), 178–87.

[9] Cheryan, M., and Rajagopalan, N. (1998). Membrane Processing of Oily Streams: Wastewater Treatment and Waste Reduction. *Journal of Membrane Science*, 151(1), 13-28.

[10] Couper J.R., Penney W.R., Fair J.R. and Walas, S.M. (2010). *Chemical Process Equipment: Selection and Design (2nd Edition)*. Elsevier Publishers, London

[11] Edison, M., John, K., Freeman, N., Mansoor, M., and Antoine, F.M-B. (2011). Cu (II) Removal from Synthetic Wastewater by Ion Exchange Process. Proceedings of the World Congress on Engineering and Computer Science, vol. II WCECS, San Francisco, USA

[12] Edomwonyi-Otu, L.C., and Adelakun, D.O. (2018). Effect of Heavy Molecular Weight Polymer on Quality of Drinking Water. *Materials Today Communications.* 15, 337–43. https://doi.org/10.1016/j.mtcomm.2018.04.004.

[13] Edomwonyi-Otu, L.C., and Angeli, P. (2015). Pressure Drop and Holdup Predictions in Horizontal Oil-Water Flows for Curved and Wavy Interfaces. *Chemical Engineering Research and Design.* 9, 55–65. https://doi.org/10.1016/j.cherd.2014.06.009.

[14] Edomwonyi-Otu, L.C., and Angeli, P. (2019). Separated Oil-Water Flows with Drag Reducing Polymers. *Experimental Thermal and Fluid Science.* 102, 467–78.

https://doi.org/10.1016/j.expthermflusci.2018.12.011.
[15] Edomwonyi-Otu, L.C., Ogabor, S.O., Idowu, O.A., and Bawa, S.G. (2013). Kinetics of Adsorption of Nickel Ion on Kankara Kaolinite. *Leonardo Electronic Journal of Practices and Technologies.* 12(22), 1–9.
[16] El-Halwagi, M.M., Gabriel, F. and Harell, D. (2003). Rigorous Graphical Targeting for Resource Conservation via Material Recycle/Reuse Networks. *Industrial & Engineering Chemistry Research*. 42(19), 4319–4328.
[17] Guerra, K., Dahm, K. and Dundorf, S. (2011). "Oil and Gas Produced Water Management and Beneficial Use in the Western United States". Science and Technology Program Report no. 157. US Department of the Interior-Bureau of Reclamation.
[18] International Petroleum Industry Environmental Conservation Association (IPIECA) (2010). Petroleum Refining Water: *Wastewater Use and Management.* 1-200.
[19] Ipsita, S.I. (2010). Wastewater Minimization Using Pinch Analysis. Bachelor of Technology (B.Tech) thesis. Under the Guidance of Prof. S. Khanam. Department of Chemical Engineering, National Institute of Technology, Rourkelamay. 2010, 1-55.
[20] Ityokumbul, M.T. (2011). Significance of Kaduna River to Kaduna Refining and Petrochemicals Complex: Some Checks and Balances, *Journal of Environmental Engineering.* 5(22), 83–98.
[21] Kaduna Refinery and Petrochemical Company (2002). Operational Annual Report. Kaduna Refinery and Petrochemical Company Ltd. Kaduna.
[22] Khezri, S.M., Lotfi, F., Tabibian, S., and Erfani, Z. (2010). Application of Water Pinch Technology for Water and Wastewater Minimization in Aluminum Anodizing Industries. *Scientific Research and Essay*. 7(2), 281–290.
[23] Lahnsteiner, J., Andrade, P., and Mittal, R.D. (2012). Recycling of Refinery and Petrochemical Effluents Employing Advanced Multi-Barrier Systems. *Water Practice and Technology*. 7(4), 12-23.
[24] Linhoff, B. (1994). Use Pinch Analysis to Knock Down Capital Costs and Emissions, *Chemical Engineering Progress.* 90(1994), 32.
[25] Mohammadnejad, S., Bidhendi, G.R.N., and Mehrdadi, N. (2016). Water Pinch Analysis in oil Refinery Using Regeneration Reuse and Recycling Consideration. *Desalination*. 265(1-3), 255–265.
[26] Nacheva, P.M. (2011). Water Management in the Petroleum Refining Industry. Mexican Institute of Water Technology, Mexico. *Water Conservation*. 105-128.
[27] NESREA (2010). National Environmental Standard and Regulation Enforcement Agency, Nigeria
[28] Peng, S., Farid, M., Ho, I., and Mahdi, K.A. (2008). Water and Wastewater Minimization in Dairy Plants Using Water Pinch Technology. *Journal of Cleaner Production.* 3(1), 43–50.
[29] Pombo, F., Alessandra M., and Alexandre S. (2009). "Technology Roadmap for Wastewater Reuse in Petroleum Refineries in Brazil". Enviromental Management in. Practice, Elzbieta Browniewicz, 65. DOI: 10.5772/20297
[30] Wang Y.P., and Smith, R. (1994). Wastewater Minimization. *Chem. Eng. Science*. 49(7), 981-100
[31] Zheng, X., Nannan Y., Xiaopeng W., Yuhong W., Linshan W., and Xiaowu, L. (2018). Adsorption Properties of Granular Activated Carbon-Supported Titanium Dioxide Particles for Dyes and Copper Ions. *Scientific Reports*. November 2017, 3–11. https://doi.org/10.1038/s41598-018-24891-1.

Reliability Assessment of Power Distribution System in the Nigerian Aviation Industry

Patience E. Orukpe,
Department of Electrical/Electronic Engineering, University of Benin, Nigeria
patience.orukpe@uniben.edu
and
Christian T. Onianwa,
Department of Electrical/Electronic Engineering, University of Benin, Nigeria
engrtojah@yahoo.com
Abdullahi Abuh,
Department of Electrical/Electronic Engineering, Federal Polytechnic Idah, Nigeria
abdullahi.idah@gmail.com
Adebisi K. Aderanti,
Electrical Department, Murtala Muhammed International Airport, Nigeria
kaadebisi@yahoo.com

Abstract: Airport terminals are the most energy intensive type of buildings due to their size, designs, activities and twenty-four hours operation. Due to recent development in airport design, incessant power failures and ineffective power distribution systems have emerged as an area of concern in the Nigerian aviation industry. To resolve these challenges, energy audit was performed by collecting energy data from three major aviation sites (airports) in Nigeria by means of data loggers, powerhouse documents and personal communication. Using collected energy data, analysis was performed to determine the reliability of these airports thereby informing the proposition of a power distribution system model for future loads. The study has contributed to the aviation industry through the development of a steady power distribution model and if implemented, the airports under evaluation would be among the leading aviation sites in the world in terms of power distribution reliability and sustainable design.

Keywords: Reliability indices, Airport, Power distribution, Power interruptions

تقييم موثوقية نظام توزيع الطاقة في صناعة الطيران النيجيرية

الملخص: مباني المطارات هي أكثر أنواع المباني كثافة في استخدام الطاقة نظرًا لحجمها وتصميمها وأنشطتها وتشغيلها لمدة ٢٤ ساعة. نظرًا للتطور الأخير في تصميم المطار، ظهرت حالات مستمرة لانقطاع التيار الكهربائي وأنظمة توزيع الطاقة غير الفعالة صارت موضوعا مثيرا للقلق في صناعة الطيران النيجيرية. لحل هذه التحديات، تم إجراء تدقيق للطاقة من خلال جمع بيانات الطاقة من ثلاثة مواقع طيران رئيسية (المطارات) في نيجيريا عن طريق مسجلات البيانات ووثائق محطة توليد الكهرباء والاتصالات الشخصية. باستخدام بيانات الطاقة التي تم جمعها، تم إجراء تحليل لتحديد موثوقية هذه المطارات وبالتالي اقتراح نموذج لنظام توزيع الطاقة للأحمال المستقبلية. ساهمت الدراسة في صناعة الطيران من خلال تطوير نموذج لتوزيع ثابت للطاقة، وفي حالة تنفيذها، ستكون المطارات قيد التقييم من بين مواقع الطيران الرائدة في العالم من حيث موثوقية توزيع الطاقة والتصميم المستدام.

1. Introduction

Power supply in Nigeria has been unreliable and unsteady, of which the aviation industries are not left out; hence the transformation currently taking place in terms of reliability, controllability and visibility. Moreover, the Nigeria government is committed to reducing overall carbon emissions by 45% by 2030 [1] and to meet renewable energy generation target of 50% demand for electricity by 2020 [2]. In order to achieve these targets set by the Nigerian government, there will be need to integrate renewable energy into the existing power system, as well as overhaul the power system infrastructure [3-5]. These functionalities will help power distribution companies in Nigeria and Distribution Network Operators (DNOs) to plan and alleviate network congestion in other to provide lower costs for electricity consumers (residential and commercial) in the airport (AP) and its environs. In this work, detailed examination of the three (3) major airports in Nigeria will be considered which include: Murtala Muhammed International Airport (MMIA) Lagos, Malam Aminu Kano International Airport (MAKIA) Kano and Nnamdi Azikwe International Airport (NAIA) Abuja. These airports have both residential and non-residential buildings. Inspection of their energy audit (facility by facility), consumption pattern, reliability of their energy distribution network configurations and energy quality of supply assessment of these airports will be closely examined so as to develop an enhanced and synchronized network distribution system for these airports [6].

In spite of the unparalleled investment in the power sector to restore the decayed distribution network facilities, the power situation in Nigeria has not improved. The gap between supply and demand keeps widening in the urban areas, while supply has not really improved in the Nigerian airports to cater for its high level of passenger traffic and operations. Ironically, experts on power generation, transmission and distribution are always giving excuses for the power crisis, which has defamed and hampered Nigeria's effort in becoming one of the most developed and industrialized nation of the world by 2020. Inadequate gas supply due to vandalism and theft, shortage of reserved water in the dams, destruction of power equipment, poor funding by private and government agencies and others, have been identified as major causes of incessant outages in these airports [7]. The Transmission Company of Nigeria (TCN) claimed that the massive load shedding/systems collapse experienced nationwide has been caused by the destruction of two major gas pipelines conveying gas to eight power generating stations in the country to run their turbines. The generating stations, which include the Olorunsogo, Egbin/AES thermal stations, Omotosho, Geregu NIPP, AFAM

IV and VI thermal power stations and Rivers State Independent Power Station were affected, resulting in drastic reduction of power supply by 1,598 MW from the current claim of 7,000MW capacity; while TCN has only been able to wheel out 5,222.3MW, the highest ever into the economy [8].

Despite all these problems and challenges affecting power system in Nigeria, management of the available energy has been of little or no concern to the power sector in Nigeria. For example in the aviation industry as at today, old underground power cables, appliances, machines, etc. that have been in operation for more than 40years are still being used in these airports. This reduces the reliability of the airport's distribution network and thereby increases its System Average Interruption Frequency Index (SAIFI) due to incessant failure and lot of energy loss along these distribution paths.

In view of the new regulation for Nigerian airports, the distribution network operator (DNO) in Nigeria, need to be able to optimize the installation and utilization of their reserve capacity in the system. This optimization is required since there is a great need to reduce cost and maximize reliability. Estimating the impact of using reserve capacity configuration, it becomes easier to determine when the risk taken is acceptable [9], so that effective planning is inevitable. Authors have suggested the use of multi-criteria dynamic optimization [10, 11], embedded generation and demand management system planning [12], system based level using energy and maximum load etc. without considering long term losses [13, 14]. The importance of considering continuous power losses in all major network planning studies based on peak load conditions, which is accepted as the basic model in calculation of power losses was presented in [15]. In [16], the author observed that efficiency of electricity systems in conjunction with cost of power losses is another important driver in reduction of network losses and should be given serious consideration. Among opportunities for loss reduction, the author described higher distribution voltages and load shifting (or system reconfiguration) as the best methods.

Series of work have been carried out on reliability assessment in electric distribution network, distribution system with distributed generation and electric transmission network [17-20]. However, the reliability assessment of the Nigerian aviation industry has not been investigated. Reliability indices used mostly are system average interruption frequency index (SAIFI), system average interruption duration index (SAIDI) and customer average interruption index (CAIDI), which will also be considered in this work. This work was necessitated based on: no available information on the energy demand, consumption pattern,

reliability of the distribution network, no chart to indicate the reliability indicators with proper network planning scheme for future loads in the three major airports.

From the foregoing, the main task of electric power systems operators is to provide their customers with a reliable and economically feasible supply of electricity [21]. As a means of providing reliable power supply, there is need to build in reserve capacity within the airports distribution substation network to cope with contingencies, increased demand for power supply at various places within the airports due to its dynamic activities and scheduled maintenance. The reserve capacity for Lagos, Abuja and Kano international airports must be designed and constructed in the most technical and economically optimal way as a viable means of having efficient and reliable power distribution at cheapest possible cost. An enhanced network planning scheme for these airport networks to improve power availability in the airport and to reduce incessant load shedding will be presented. Hence, commonly used method of reliability analysis will be adopted in this work.

2. Methodology

The procedure used in order to execute this work is explained in this section.

The probability distributions instead of simple average techniques and standard deviations will be used to evaluate the effect on reliability indicators (e.g. **System Average Interruption Duration Index (SAIDI), Consumer Average Interruption Duration Index (CAIDI),** and **System Average Interruption Frequency Index (SAIFI)** of the different cases to be examined. Different cases were examined for Lagos, Abuja and Kano airports through SAIFI, SAIDI, CAIDI and frequency of interruptions in these airports.

Tables would be developed for each case for the data obtained in 2016 for the airports under consideration and these tables would be used to form charts for each case using reliability indices. These tables would also be used to plot graphs of reliability indices for each airport, identify which, what and why, then what can be done to improve the systems reliability for short time and for future expansion. For the future expansion, new airport would be modeled for Lagos and recommendation will be made in relation to the other two airports.

2.1 Data Collection and Collation

Energy survey and data gathering from the three (3) major airports in Nigeria using both primary and secondary data, energy audit of these airport so as to determine their load capacity, quality of supply assessment, energy consumption pattern and energy performance score for these airport as means of determining the airport that has the most reliable distribution

network. The use of value of consumers' reliability indices to develop a more efficient power system networks for these airports was carried out.

From the survey carried out on buildings and offices in the airports, data collected from electrical system control rooms, energy audit and manufacturers' label on equipment were collated and presented as in Tables 1, 2, 3 and 5. Furthermore, graphs were developed as presented in Figures 1, 2, 3 and 4 for interruptions and reliability indices for the airports under consideration.

2.2 Reliability Indices

Consumers' based reliability indices considered in this work are SAIDI, SAIFI and CAIDI. SAIDI is defined as the total duration of sustained interruptions in a year per total number of consumers as given in (1) [20,21] and its unit is 'minutes'; SAIFI is defined as the total number of sustained interruptions in a year per total number of consumers as given in (2) [20] and its unit is 'interruptions per consumers'; CAIDI is defined as the total duration of sustained interruptions in a year per total number of interruptions and it is also the ratio of SAIDI to SAIFI and is given by the expression in (3) [20]. The unit of CAIDI is 'minutes'.

$$SAIDI = \frac{Sum\ of\ consumers\ interruptions\ duration}{Total\ number\ of\ consumers\ served} \quad (1)$$

$$SAIFI = \frac{Total\ number\ of\ consumers\ interrupted}{Total\ number\ of\ consumers\ served} \quad (2)$$

$$CAIDI = \frac{Sum\ of\ consumers\ interruptions\ duration}{Total\ number\ of\ consumers\ interrupted} = \frac{SAIDI}{SAIFI} \quad (3)$$

In this work, SAIFI, SAIDI and CAIDI are used since they are the reliability indices related to consumer supply calculated using annual field data.

2.3 Reliability

Reliability in power system is the ability of the system to satisfy the consumer demand. In this work reliability will be calculated using (4) [22] and it will be used to validate the reliability of the three airports under consideration.

$$Reliability = \frac{(Total\ minutes\ in\ year - Total\ minutes\ without\ supply)\times 100}{Total\ minutes\ in\ year}$$

$$= \frac{\left(365\ days\ \times \frac{24\ Hours}{day}\ \times \frac{60 minutes}{Hour} - Duration\ of\ outages(minute)\right)\times\ 100}{\left(365\ days\ \times \frac{24\ Hours}{day} \times\ 60 \frac{minutes}{Hour}\right)} \quad (4)$$

3. Results and Discussion

The Lagos airport has two supplies from ISOLO and EJIGBO power substations. The ISOLO power station consists of a 132/33kV, 1 x 60MVA step down On Load Transformer Changer (OLTC) transformers, while the Ejigbo substation has 132/33kV, 2 x 100MVA step down OLTC transformers. The secondary of these transformers were fed to Lagos airport via a 33kV panel having a breaker capacity of 1250A. This energy is first received by a network of UG cables that were fed to copper bus coupler, then connected to 33kV/11kV step down transformer for further handling of the energy. Evaluated data for the year 2016 showing interruptions frequency and duration is shown in Table 1 for Lagos, while Table 2 and Table 3 show that for Abuja and Kano respectively.

Table 1: Lagos unplanned and planned long consumers interruptions

From energy data recordings of 12,236 consumers Voltage level (99.86% at 0.415kV)					
Months in 2016	Consumers Interruption		SAIFI (Interruption /Consumers)	SAIDI (Minutes)	CAIDI (Minutes)
	No. Interruption	Duration (minutes)			
January	368	52,293	0.03	4.27	142.10
February	704	51,036	0.06	4.17	72.49
March	822	59,803	0.07	4.89	72.75
April	1,014	62,651	0.08	5.12	61.79
May	1,172	81,483	0.10	6.66	69.52
June	1,494	75,627	0.12	6.18	50.62
July	225	41,924	0.02	3.42	186.33
August	364	47,335	0.03	3.87	130.04
September	322	44,211	0.03	3.61	137.30
October	469	55,027	0.04	4.50	117.33
November	544	53,312	0.04	4.36	98.00
December	562	58,998	0.05	4.82	104.98
Total	8,060	683,700	0.67	55.87	1243.25

From Table 1, the highest number of interruption occurred in the month of June with the least consumers' interruption duration index, while the highest period of interruption in minutes occurred in the month of May. It is observed that, the number of interruption is not directly proportional to the duration of interruption. Since SAIFI and SAIDI are functions of interruption, it is observed that, the higher the number of interruption, the higher the magnitude of SAIFI and SAIDI.

Table 2: Abuja unplanned and planned long consumers interruptions

From energy data recordings of 4,589 Consumers Voltage level (99.53% at 0.415kV)					
Months in 2016	Consumers Interruption		SAIFI (Interruption /Consumers)	SAIDI (Minutes)	CAIDI (Minutes)
	No. Interruption	Duration (minutes)			
January	352	37,814	0.07	8.24	107.43
February	372	50,968	0.08	11.11	137.01
March	411	63,067	0.09	13.74	121.28
April	351	42,571	0.08	9.28	153.45
May	241	39,332	0.05	8.57	163.20
June	105	34,668	0.02	7.55	330.17
July	140	36,746	0.03	8.01	262.47
August	237	39,668	0.05	8.64	167.38
September	229	40,739	0.05	8.88	177.90
October	276	39,893	0.06	8.69	144.53
November	276	42,738	0.06	9.31	154.85
December	350	47,835	0.08	10.42	136.67
Total	3,340	516,136	0.72	112.44	2056.34

From Table 2, the highest number of interruption occurred in the month of March with the highest period of interruption duration in minutes, while the least consumers' interruption duration index is in January. It is also observed that, the number of interruption is not directly proportional to the duration of interruption. Similarly, SAIFI and SAIDI increases relative to the number of interruptions.

Table 3: Kano unplanned and planned long consumers interruptions

From energy data recordings of 3,019 Consumers Voltage level (99.91% at 0.415kV)					
Months in 2016	Consumers Interruption		SAIFI (Interruption/ Consumers)	SAIDI (Minutes)	CAIDI (Minutes)
	No. Interruption	Duration (minutes)			
January	163	10,414	0.05	3.45	63.89
February	167	15,482	0.06	5.13	92.71
March	172	17,336	0.06	5.74	100.79
April	148	15,435	0.05	5.11	104.29
May	142	14,996	0.05	4.97	105.61
June	113	10,325	0.04	3.42	91.37
July	122	10,052	0.04	3.33	82.39
August	128	12,528	0.04	4.15	97.88
September	134	11,047	0.04	3.66	82.44
October	149	12,974	0.05	4.30	87.07
November	153	13,222	0.05	4.38	86.41
December	157	13,749	0.05	4.55	87.57
Total	1,748	157,560	0.58	52.19	1,082.42

From Table 3, the highest number of interruption occurred in the month of March with the highest period of interruption duration in minutes, while the least consumers' interruption duration index is in January. It is also observed that, the number of interruption is not directly proportional to the duration of interruption. In a similar vein, SAIFI and SAIDI increase with increasing number of interruption. From Tables 1 to 3, it can be seen that Lagos, Abuja and Kano airports followed the same pattern in terms of interruption, duration and reliability indices.

Plotting the graph of outages in Lagos, Abuja and Kano against twelve calendar months in the year 2016, the number of interruptions, SAIFI, SAIDI and CAIDI are shown in Figures 1 - 4.

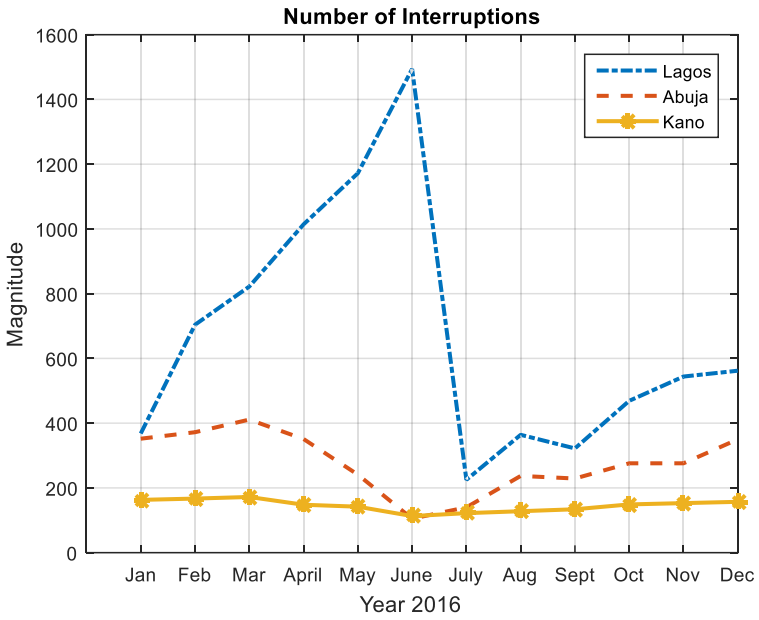


Figure 1: Lagos, Abuja and Kano graph of number of interruption.

Figure 1 shows behaviors of Lagos, Abuja and Kano interruptions in the year 2016. Kano has peak interruption in the month of March which is less than 200 times. Abuja has peak interruption in the month of March too but a little bit above 400 times, while Lagos with the least interruption in the month of July which is a little above 200, but has highest peak in the month of June and frequency of about 1500 times indicating instability in the distribution system.

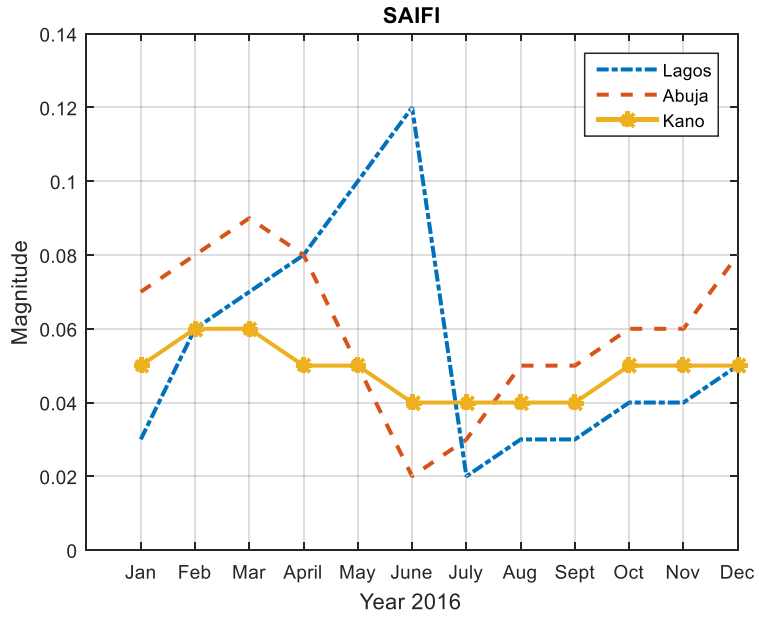


Figure 2: Lagos, Abuja and Kano SAIFI graph.

Looking at the behaviors in Figure 2, Kano airport has the most improved SAIFI oscillating between 0.04 and 0.06 which is the ratio of the number of interruption to the number of consumers. This means that, Kano airport is the most stable airport among the three airports followed by Abuja airport. Lagos has the highest number of interruptions indicating an unstable power distribution system.

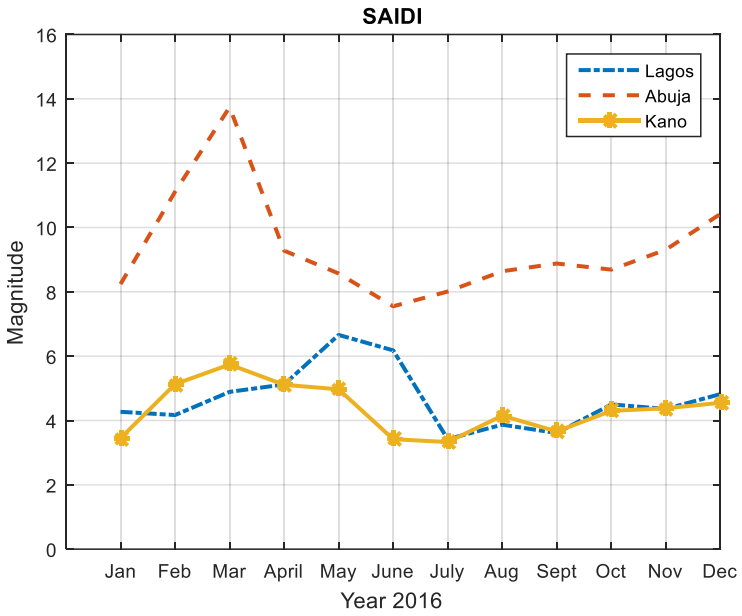


Figure 3: Lagos, Abuja and Kano SAIDI graph

In Figure 3, Kano airport has the least SAIDI followed by Abuja, while Lagos has the highest system of interruption in the year 2016. Lagos still remains unstable with this reliability index, due to the high number of consumers when compared with the other airports.

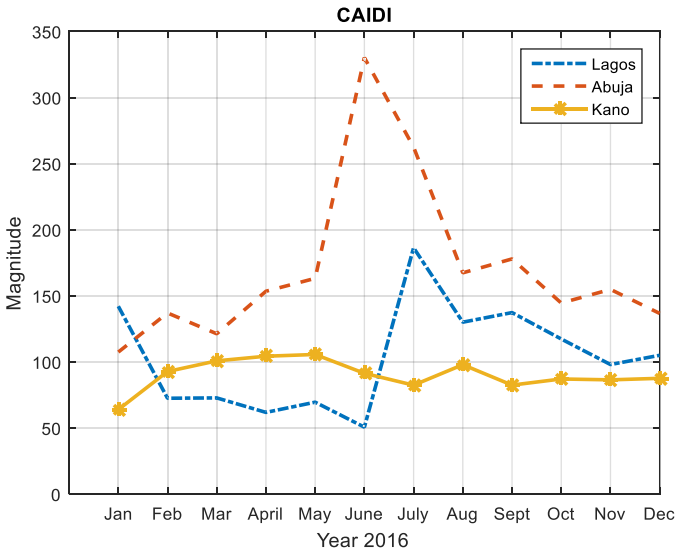


Figure 4: Lagos, Abuja and Kano CAIDI graph.

Although the CAIDI graph behaviors found Abuja consumers interrupted the most with highest index of about 340. However, from the data presented in Table 1, Lagos has the highest number of interruptions when compared with Abuja, still making it the most unstable airport due to the large number of consumers in it. Kano remains the most stable airport among the three airports as depicted in Figure 4.

To validate the reliability of the airports, we use the mathematical equation in (4). The reliability for Lagos, Abuja and Kano is calculated as follows:

a. Lagos airport having 683,700minutes of outages, has a reliability of:

$$Reliability_{Lagos} = \frac{\left(365\ days\ \times \frac{24\ Hours}{day}\ \times \frac{60minutes}{Hour} - 683{,}700\right) \times\ 100}{\left(365\ days\ \times \frac{24\ Hours}{day} \times\ 60\frac{minutes}{Hour}\right)}$$

$$= \frac{(525{,}600 - 683{,}700) \times\ 100}{(525{,}600)} = -30\%$$

b. Abuja airport having 516,136minutes of outages, has a reliability of:

$$Reliability_{Abuja} = \frac{\left(365\ days\ \times \frac{24\ Hours}{day}\ \times \frac{60minutes}{Hour} -\ 516{,}136\right) \times\ 100}{\left(365\ days\ \times \frac{24\ Hours}{day} \times\ 60\frac{minutes}{Hour}\right)}$$

$$= \frac{(525{,}600 - 516{,}136) \times\ 100}{(525{,}600)} = 1.8\%$$

c. Kano airport having 157,560 minutes of outages, has a reliability of:

$$Reliability_{Kano} = \frac{\left(365\ days\ \times \frac{24\ Hours}{day}\ \times \frac{60minutes}{Hour} -\ 157{,}560\right) \times\ 100}{\left(365\ days\ \times \frac{24\ Hours}{day} \times\ 60\frac{minutes}{Hour}\right)}$$

$$= \frac{(525{,}600 - 157{,}560) \times\ 100}{(525{,}600)} = 70\%$$

The reliability indices and the reliability for Lagos, Abuja and Kano are presented in Table 4.

Table 4: Summary of reliability indices for Lagos, Abuja and Kano airports

Year, 2016	No. of Consumers	No. of interruption	Duration of interruption (minutes)	SAIFI	SAIDI	CAIDI	Calculated Reliability
LAGOS	12,236	8,060	683,700	0.67	55.87	1,243.25	-30%
ABUJA	4,589	3,340	516,136	0.72	112.44	2,056.34	1.8%
KANO	3,019	1,748	157,560	0.58	52.19	1,082.42	70%

From all indication, considering Table 4, Lagos airport takes second position in the table in terms of SAIFI, SAIDI and CAIDI but needs more attention considering the fact that it is the most unstable, as well as the most patronized airport among the three with the highest number of consumers and having the negative index of (-30%) calculated reliability (the calculated reliability validates the fact that Lagos is unstable/unreliable). Attention would be on Lagos airport in order to relate causes of interruption both planned and unplanned as they affect planning, reliability and quality of power distribution systems.

4. Projected Lagos Airport Consumers

The major causes of power outage in Lagos aside from power failure from national grid that are recurrent, which results in its being the most unstable among the airports despite the fact that the airport is the most patronized, would be singled out for year 2017 with a projected increased number of consumers to 13,000. Table 5 is the Lagos 2017 isolated causes of interruption, number of interruption (274), duration (10,120 minutes) and reliability indices (SAIFI, SAIDI and CAIDI); while Table 6 shows differences between Lagos 2016 and selected 2017 reliability parameters.

Table 5: Lagos power house internal planned and unplanned interruption (interruption causes)

Lagos	From energy data recordings of 13,000 Consumers Voltage level (99.86% at 11kV)							
Year 2017	Interruption Causes			Customer interruption		SAIFI (Interruption /Consumers)	SAIDI (Minutes)	CAIDI (Minutes)
	Cable	Load shedding	Water	No.	Duration (minutes)			
January	14	8	0	22	1,603	0.0017	0.1233	72.86
February	7	13	0	20	660	0.0015	0.0508	33.00
March	4	40	0	44	1,273	0.0034	0.0979	28.93
April	2	15	0	17	588	0.0013	0.0452	34.59
May	8	8	3	19	897	0.0014	0.0690	47.21
June	2	0	8	10	512	0.0008	0.0394	51.20
July	4	0	13	17	562	0.0013	0.0432	33.06
August	0	3	0	3	9	0.0002	0.0006	03.00
September	2	0	14	16	1,147	0.0012	0.0882	71.69
October	9	20	1	30	1,213	0.0023	0.0933	40.43
November	10	25	0	35	1,505	0.0027	0.1158	43.00
December	11	30	0	41	151	0.0032	0.0116	03.68
Total	73	162	39	274	10,120	0.0296	0.5080	469.65

Table 6: Summary of Lagos 2016 and 2017 Parameters

LAGOS	No. Of Consumers	No. of interruption	Duration of interruption (minutes)	SAIFI (Interruption/ Consumers)	SAIDI (Minutes)	CAIDI (Minutes)
Year, 2016	12,236	8,060	683,700	0.67	55.87	1,243.25
Year, 2017	13,000	274	10,120	0.0296	0.5080	469.65

It can be seen in Table 1 for year 2016 that, the total number of approximated consumers connected to the 11kV feeders is 12,236 with total number of interruption of 8,060 having a duration of 683,700minutes, while that of selected causes has 13,000 consumers with total

number of interruption of 274 having duration of 10,120minutes. It is observed that, there is reduction in the number of outages and interruption despite increment in the number of consumers in 2017 even though it was selected compared to 2016. It can be seen from the Tables and Figures that the chart follows the same pattern, although there is abnormality in the data of Table 1 where there is high number of interruption between May and July with peak in June. This was as a result of gas pipelines vandalism when the change of Nigeria Government took place. From Table 5, the factors responsible for major outages between six months of October and March are typical of the characteristic dry season of those periods. The transmitted power to the injection stations is always on the low owing to the limited energy generated as a result of low volume of water at the generation end, even though the energy generated through gas turbines would be distributed to transmission lines which would not be enough. If no power is transmitted to injection substation, there would be no power supply to the end users.

In the dry season, consumers tend to use more of energy consuming equipment like air conditioning systems as a result of more passenger traffic and the heavy heat during the period. The air conditioners inspected both at the residential and commercial areas are very powerful thereby consuming more energy, hence resulting in overloading of the system. In addition to these air conditioners energy consumptions, newly built hanger, commercial centers like car park and terminals are the contributory factors to the system overloading. Since there are only two 8MVA transformers, the new developments are adding to the accumulated load on a daily basis; and to avoid system collapse since combined capacity of the two transformers cannot power more than 13MW, load shedding is unavoidable. On an hourly basis, the total load is usually monitored and documented in the power house log book. When the load is above 12.5MW, load shedding done by mechanically opening/de-energizing the breakers of the feeders supplying some substations is usually carried out because demand exceeds power supply.

5. Remodeling of the Airports with Reserved Capacity

Following directly from the identified challenges, the Lagos airport was remodeled as presented in Figure 5 to help improve the network quality and reliability.

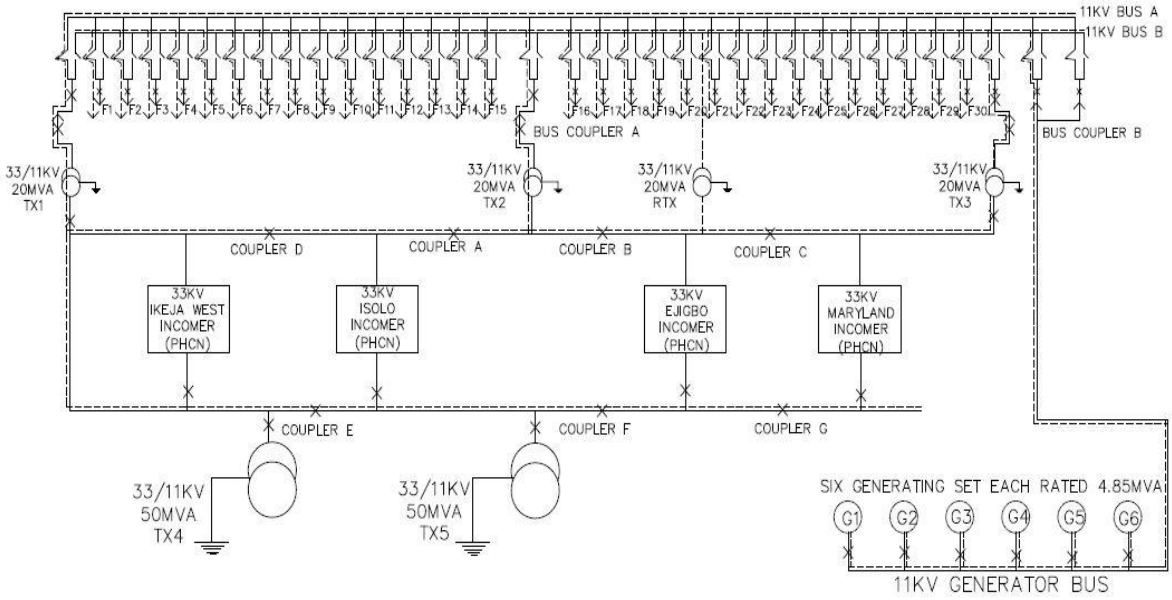


Figure 5: Lagos remodeled airport.

Figure 5 presents the remodeled Lagos airport for the forecasted load. The remodeled airport is a modification of the existing airport to accommodate current excess power consumption and to cater for future load for the next decade. When power equipment were evaluated across airports, Lagos was seen to be the busiest among all Nigerian airports, as a result, load predictions were made for this airport. Based on the load prediction, reliability and quality of power distribution were considered in remodeling the system. The airport has more additional power sources which are IKEJA WEST and MARYLAND power substations to complement existing EJIGBO and ISOLO power sources. The additional sources will increase the system quality of power supply resulting in higher reliability. Only a single source can be used at a time, while the remaining sources are redundant. From existing Lagos airport diagram, the two functional transformers having total output of 15MW is replaced with three numbers of 20MVA transformers which would produce a total output of 48MW. This shows that even with 40MW forecasted, there is 8MW tolerance to cater for unplanned circumstances.

Power supply from ISOLO and EJIGBO substation was via UG 2x33kV, 1x240mm^2 XLPE armored cable for ISOLO and 3x240mm^2 for EJIGBO, both lengths of approx. 100km in duct. Cables with dry season continuous rating of 437A (~8MVA) each are presently energized at 11kV. This same rating of armored XLPE cable would be used for newly introduced Ikeja West and Maryland in Lagos. Aside from the three numbers of 20MVA proposed in the remodeled airport in Lagos, one 20MVA is placed as redundant should there be fault in any of the three numbers of 20MVA as represented with dotted transformer and labeled RTX. Two numbers of

50MVA are placed to serve as transformer reserve capacity for future expansion or to be restored to the system in case we have worst scenarios when two numbers of 20MVA collapsed or are not synchronize-able. The current system has 13 functional distribution feeders feeding different sections of the airport in Lagos. The loads on these feeders are much and complex as a result of many organizations, companies along the same line that are looped; hence the need for more flexible system with more feeders to serve areas looped together under the same feeder currently. The feeders F1, F2, …F13 as they are currently constituted in Lagos, is modeled to 30 feeders F1, F2, …F30 with a redundant cable line along each feeder to serve as backup in case there is cable fault along the line. All the cable lines duplicated in dotted line are redundant cable lines to prevent power outage, reduce the down time and also reduce restoring time since consumers are ready to pay for their consumptions. The case of Abuja and Kano airport as cross examined, the two numbers of 2MVA transformers used in Abuja delivering a total output of 4MW is not enough to accommodate for the network and its reserve capacity. By audit, the total load that will conveniently feed the whole airport is estimated at 7MW, while the case of Kano having two numbers of 800kVA transformers will not adequately feed its 3.2MW load capacity. Therefore, there is need for upgrade in these airports to accommodate more load and reserve capacity. The existing distribution network can be modified to accommodate more loads. The Abuja 2.5MVA can be replaced with two numbers of 8MVA and 800kVA can be replaced with equivalent 2.5MVA transformers in Kano airport. The power transformers used in the existing airports are presented in Table 7.

Table 7: Existing airport transformers.

Airports	Transformers	Power output	Energy Audit Peak Load
Kano	2 × 800kVA	3.2MW	1.39MW
Abuja	2 × 2MVA	4MW	2.87MW
Lagos	2 × 8MVA	12.8MW	12.6MW

When the capacity loads of these airport are being increased, it will be possible to design a more reliable distribution network configuration for these airports.

Table 8: Remodeled airport transformers with reserve capacity.

Airports	Transformers	Power output	Energy Audit Peak Load	Reserve capacity
Kano	2 × 2.5MVA	3.7MW	5MW	-
Abuja	2 × 8MVA	12.8MW	8MW	-
Lagos	3 × 20MVA	48MW	15MW	1x20MVA - 2x50MVA

Table 8 shows the summary of transformers modeled for 2025 forecasted load for Lagos, Abuja and Kano airports (to cater for the airports upgrade and increased number of consumers). Since there is lesser activities in MAKIA Kano airport and it is the most reliable airport among the three airport under survey, there is need for little power distribution upgrade. The Abuja airport present load is used to forecast and model future power consumption which is pegged at 10MW. This will accommodate future load with reserve capacity. Lagos forecasted load for 2025 is pegged at 40MW. As a result of Lagos airport importance, being the commercial centre of the nation, the newly remodeled airport has redundant plant of 20MVA with reserve capacity 2x50MVA. From the foregoing, increased load of 40MW, 10MW and 3.7MW is projected for Lagos, Abuja and Kano airports respectively in the future to cater for airport growth in terms of commercial and industrial developments (see Table 8).

This robust airport proposal is to accommodate various unplanned loads and to ensure network security through this planning, reliability and quality. If this remodeled airport is implemented, Lagos would be one of the leading airports in the world in terms of reliability and quality of power distribution resulting in system security.

6. Conclusion

In this work, reliability assessment of power distribution in three major airports was considered. From the survey carried out, data was collected and collated for the airports for the first time, which can be used by other researchers. This work has conducted energy audit in selected airports in the aviation industry of Nigeria and has identified some grey areas in the distribution network indicative of poor reliability of the system. From the reliability indices (SAIFI, SAIDI and CAIDI) obtained from the data and the calculated reliability, it was observed that Lagos airport is the most unstable/unreliable due to the large traffic experienced (most patronized) and the highest number of consumers in the airport. An expanded model with redundant links has been proposed for reliable and sustainable energy in the aviation sector by running a predictive model for future energy expansion up till 2025.

7. References

[1] NINDC, Nigeria's Intended Nationally Determined Contribution, Paris climate change summit, http://www4.unfccc.int/ndcregistry/PublishedDocuments/Nigeria%20First/Approved%20Nigeria%27s%20INDC_271115.pdf, 2015 (accessed 20 May 2017)

[2] I. Chiemeka, T. Chineke, Evaluating the global solar energy potential at Uturu, Nigeria, International Journal of Physical Sciences, 4(3) (2009) 115 – 119.

[3] NERC, Potential reliability impacts of emerging flexible resources, North American Electric Reliability Corporation, Tech. Rep., 2010.

[4] E. Lannoye, D. Flynn, M. O'Malley, Evaluation of power system flexibility, IEEE Trans. on Sys., 2 (2012) 922–931.

[5] V. C. Gungor, B. Lu, G. P. Hancke, Opportunities and challenges of wireless sensor networks in smart grid, IEEE Trans. on Ind. Electron., 57(10) (2010) 3557-3564.
[6] P. P. Xu, H. W. Chan, K. Qian, Critical success factors (CSFs) for sustainable building energy efficiency retrofit (BEER) using energy performance contracting (EPC) for Hotel buildings in China, Energy Policy, 39 (2011) 7389-7398.
[7] R. Fashola, Nigeria Ministry of Power, Power supplies worsen as TCN, NNPC records system collapse, Nigeria Daily Post, May 28, 2017.
[8] R. Fashola, Nigeria transmits 5,222.3MW electricity highest ever, Nigeria Premium Times, January 26, 2018.
[9] G. Purdy, ISO 31000:2009 setting a new standard for risk management, Risk analysis; An official publication of the Society for Risk Analysis, 30(6) (2010) 881-886.
[10] V. Neimane, G. Andersson, Distribution networks reinforcement planning: a dynamic multi-criteria approach, In Proceedings of International Conference on Electric Power Engineering, Power Tech Budapest 1999.
[11] V. Neimane, On development planning of electricity distribution networks, Doctoral Thesis, Electrical Systems, KTH, 2001.
[12] Sinclair Knight Merz's (SKM) Report for the Austalian energy market commission, (2005) 25-40.
[13] ACIL Tasman, System energy guidelines, common forecasting methodology prepared for Ergon Energy and Energex, Brisbane, 2010.
[14] ACIL Tasman, System maximum demand guidelines, common forecasting methodology prepared for Ergon Energy and Energex, Brisbane, 2010.
[15] N. Ivosevic, Elektroenergetika, Osnovni principi i energetsko-ekonomske Informacije, (Power Engineering, The Basic Principles and Energy and Economic Information), Institut za Naucno-Tehnicku Dokumentaciju i Informacije (Institute for Scientific and Technical Documentation and Information), Beograd (on Serbian), 1970.
[16] G. J. Anders, Rating of Electric Power Cables, McGraw Hill, IEEE Press, New York, 1997.
[17] P. U. Okorie, U. O. Aliyu, B. Jimoh, S. M. Sani, Reliability indices of electric distribution network system assessment, 3(1) (2015) 1-6.
[18] M. M. A. Al-Maghalseh, Evaluating the reliability worth indices of electrical medium voltage network: case study, Procedia Computer Science, 130 (2018) 744-752.
[19] G. Chen, F. Zhang, D. You, Y. Wang, G. Lu, Q. Zou, H. Liu, J. Qian, H. Xu, Reliability evaluation of distribution system with distributed generation, IOP Conf. Series: Materials Science and Engineering, 224 (2017), 1-9.
[20] J. O. Aibangbee, I. N. Chukwuemeka, Reliability assessment of APO 132/33kV electric transmission substation Abuja, American Journal of Engineering Research, 6(8) (2017) 170-176.
[21] R. Billinton, R. Allan, Reliability Evaluation of Power Systems, New York: Plenum Press, 1996.
[22] G. S. Ivonavich, Modern distribution network planning and application to development of electricity infrastructure, Master's Thesis, University of Southern Queensland, Faculty of Engineering and Surveying, 2010.

Energetic and Exergetic Investigation of Diesel - Brayton and Organic Rankine Combined Cycle

M N Khan,
Department of Mechanical and Industrial Engineering, College of Engineering,
Majmaah University, KSA
mn.khan@mu.edu.sa

Abstract: The amount of energy lost through the exhaust gases from diesel as well as gas turbine power plant is much enough to enhance the performance of the power system as well as to operate another secondary power generation system. If this energy is not to be used, it will results in global warming and enhance the local air pollution problem. Present study proposes a new cycle to utilizes exhaust gas energy from diesel engines and gas turbine power plants to optimize the performance of the combined cycle. In the proposed system, the waste energy from the exhaust gas of the diesel power plant is to used to enhance the thermal efficiency of the gas turbine cycle (GTC), and the waste energy from the exhaust gases of the GTC is to utilized to operate an organic Rankine cycle (ORC). To investigate the overall performance of the proposed system, the pressure ratio and turbine inlet temperature (TIT) of GTC are used as variables. The result shows that the gain in thermal efficiency and work output of the proposed system is about 15% to 43% and 13% to 20.4% as compared to diesel plus simple GTC system when TIT increases from 900K to 1200K. The exhaust gases exergy loss from the diesel system and simple GTC is much high as compared to the proposed-system. The total exhaust gases exergy loss from the proposed-system is decreased by 3.5 to 1.35 times as compared to the total exhaust gases exergy loss from diesel cycle and simple GTC.

Keywords: Work output, Thermal efficiency, SFC, Exergy loss, Diesel engine, GTC, ORC

بحث نشط ودقيق للديزل - دورة برايتون ورانكين العضوية المركبة

الملخص: إن كمية الطاقة المفقودة من خلال غازات العادم من الديزل وكذلك محطة توليد الطاقة التوربينية الغازية كافية لتحسين أداء نظام الطاقة وكذلك لتشغيل نظام ثانوي آخر لتوليد الطاقة. إذا لم يتم استخدام هذه الطاقة، فسيؤدي ذلك إلى ارتفاع درجة حرارة الأرض وتعزيز مشكلة تلوث الهواء المحلي. تقترح الدراسة الحالية دورة جديدة لاستخدام طاقة غاز العادم من محركات الديزل ومحطات الطاقة التوربينية الغازية لتحسين أداء الدورة المركبة. في النظام المقترح، يتم استخدام الطاقة المهدرة من غاز العادم لمحطة توليد الطاقة بالديزل لتعزيز الكفاءة الحرارية لدورة التوربينات الغازية (GTC)، ويتم استخدام الطاقة المهدرة من غازات العادم في GTC للعمل دورة رانكين عضوية (ORC). لاستكشاف الأداء العام للنظام المقترح، يتم استخدام نسبة الضغط ودرجة حرارة مدخل التوربينات (TIT) لـ GTC كمتغيرات. تظهر النتيجة أن الكسب في الكفاءة الحرارية وإخراج العمل للنظام المقترح هو حوالي ١٥٪ إلى ٤٣٪ و١٣٪ إلى ٢٠.٤٪ مقارنة بالديزل بالإضافة إلى نظام GTC البسيط عندما يزيد TIT من ٩٠٠ ألف إلى ١٢٠٠ ألف. إن فقدان الطاقة الناتجة عن غازات العادم من نظام الديزل وGTC البسيط مرتفع للغاية مقارنة بالنظام المقترح. تم تقليل إجمالي فقدان الطاقة الناتجة عن غازات العادم من النظام المقترح بمقدار ٣.٥ إلى ١.٣٥ مرة مقارنة بإجمالي فقدان الطاقة الناتجة عن غازات العادم من دورة الديزل وGTC البسيط.

1. Introduction

Now a day, the domestic, as well as industrials demand for electrical energy worldwide, is mainly covered by the massive capacity power plant. Due to the continuous increase of fuel prices at the international level and continuous power demand, a simple gas turbine cycle (GTC) based power plant and diesel power plant is not sufficient to fulfill the demand for electrical power and price of power generation. Generally, the temperature of exhaust gases from the GTC plant and diesel plant is very high that results in global warming as well as increase air pollution. The energy carried out by the exhaust gases from large capacity GTC and diesel power plant is much enough that is capable to run another cycle which may be GTC, Rankine cycle (RC), and organic Rankine cycle (ORC). Whatever is the cycle, in general such cycles are known as the bottoming cycle. The results of the operation of the bottoming cycle by the exhaust of the topping cycle (GTC or Diesel power plant), helps to reduce global warming and air pollution [1–6]. Power generation from the exhaust of the topping cycle gives the concept of a combined cycle power plant (CCPP). From last decade, several researchers show that some major factors affect the overall performance of CCPP like turbine inlet temperature (*TIT*) [7–9], air compressor inlet temperature and density [10–13], and last but not the least is air compressor pressure ratio [8,14]. Hüseyin et al. [15] parametrically investigate the organic Rankine cycle (ORC) in the case of supercritical and subcritical operated by the energy of exhaust gases of power engine and biogas-fueled combined cycle. Results show that in comparison to subcritical ORC and supercritical ORC, supercritical ORC is superior to subcritical ORC in terms of energy and exergy efficiency, and net power. Chen et al. [16] thermodynamically investigated the open combined Brayton and two parallel inverse Brayton cycles. Results show that by regulating the compressor inlet pressure of the bottom cycles, the proposed model performance was improved. Khan [17] investigate the performance of air bottoming combined cycle and regenerative gas turbine cycle operated by the partial amount of exhaust gasses from the gas turbine. This study presents the unique technique to compare the performance of these two cycles and prove that for thermal efficiency and exhaust gasses exergy loss by regenerative GTC is much better as compared to the air bottoming cycle but for net power output air bottoming cycle is better than regenerative GTC. In another study, Khan and Tlili [9] investigated the importance of heat exchanger in regenerative topping gas turbine and air bottoming cycle connecting through the bypass valve. The study proves that by proper use of heat exchanger and bypass valve work output of the combined cycle increases from 13.5 to 45% and combined cycle efficiency increases from 15% to 31%. Mishra and kumar [18]

proposed organic bottoming cycle operated by the energy carried out by the exhaust gases from the gas turbine cycle. The study was conducted for two different fluids that is R123 and R245fa to evaluate the overall performance of combined system. Results show that the thermal efficiency of combined cycle in case of R123 is more as compared to R245fa but the work output of combined cycle for R123 is more as compared to R245fa. Galindo et al. [19] theoretically investigated the Brayton cycle operated by the exhaust of a passenger car. In this study, energy recovery from the exhaust from a 2-L turbocharged gasoline engine was suggested. The result shows that the suggested method increased engine efficiency by up to 15%. Zhu et al. [20] theoretically investigate the energy and exergy analyses of the bottoming Rankine cycle run by the heat carried by the exhaust gases from the engine. Results show that out of five selected fluids for the engine, R113 and ethanol reflect the best thermodynamic performance over the complete exhaust gasses temperature range. Also, the main factor that influences the system performance is superheating temperature, working fluid properties, and evaporating pressure.

The above literature reveals that many factors affect the performance of the combined cycle power plant (CCPP). A several number of techniques have been suggested by several researchers to boost the performance of the combined cycle power plant (CCPP) [21–26]. In the proposed system, the exhaust of the diesel engine is used to preheat the air to the combustion chamber of GTC, which results in a decrease of fuel supply for attending the required *TIT*. Also, the exhaust of the GTC used to operate the ORG where working fluid is R134a under the present study.

The novelty of the present study is that the work output and thermal efficiency of the proposed system is much effective as related to Simple GTC and diesel cycle. The exhaust gases exergy losses from the proposed system is much less as compared to the exhaust gases exergy losses from the diesel cycle and simple GTC. This means that the proposed system is very efficient to reduce the exhaust gases exergy losses from the diesel cycle as well as from Simple GTC. The application of the present study is directly applicable where the temperature of exhaust from the diesel power plant is very high and the energy carried out by the exhaust gases from the diesel power plant is enough to preheat the compressed air of the GTC power plant. The one of the major challenge of the present system is that the proposed system is more complicated in design and operation as compared to the Diesel power plant and simple GTC power plant. The graphical presentation of energy and exergy results are helpful for the readers to understand the uniqueness of the proposed system.

2. Cycle description

Figure 1 and fig. 2 shows the schematic and T-s diagram of the proposed system. Air from the environment at atmospheric temperature and pressure T_1 and P_1 enter the diesel engine. The exhaust gases from the diesel engine at temperature T_4 and pressure P_4 leaves to enters the environment via a heat exchanger (*H.E.*) where it heats the compressed air from the air compressor of the GTC plant. This results in a temperature decrease of exhaust gases of diesel engine from T_4 to T_5 and the compressed air temperature increased from T_7 to T_8. The net work output, mass flow rate of flue gases and thermal efficiency of the diesel engine are given by the equation (4), equation (8), and equation (9) respectively.

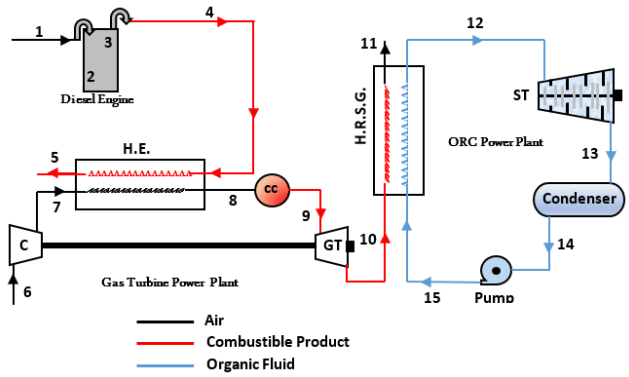


Figure 1. Schematic diagram of the proposed system

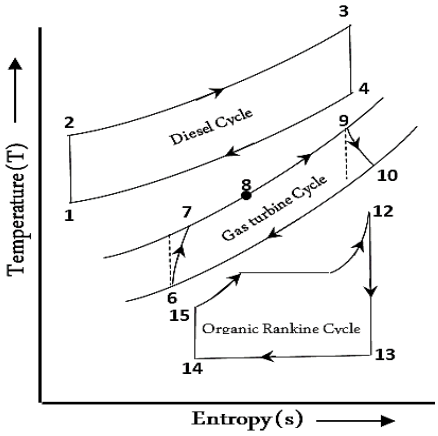


Figure 2. T-s diagram of the proposed system

In GTC plant, air from the environment at atmospheric temperature and pressure T_1 and P_1 enters the air compressor and leaves at temperature T_2 and pressure P_2. The compressed air at temperature T_7 and pressure $P_7=P_8$ enters the combustion chamber via the *H.E.* where it mixes with fuel (Natural gas) resulting in the rise of combustion products temperature to T_9 (*TIT*). The combustible product enters the heat recovery steam generator (HRSG) at temperature T_{10} and pressure P_{10} via a gas turbine. The exhaust gases from the gas turbine leave to the environment via HRSG exchange which heats to the refrigerant (R134a) of the organic Rankine cycle (ORC). The governing equations for work output, thermal efficiency, SFC, etc. for GTC are listed below from equation 10 to equation 21. In ORG, refrigerant at pressure P_{15} enters the HRSG where it gains heat from the exhaust gases of the GTC plant. The refrigerant at pressure P_{15} expands to pressure P_{13} in the turbine. The governing equations for ORG are listed below from equation 22 to equation 25. The major assumptions considered for the analysis of the proposed system are itemized in table 1.

Table 1. List of supposed constraints and variables [11,27,28].

Variables		
Turbine inlet temperature of GTC	*TIT (K)*	900 to 1200
Pressure ratio of GTC	r_p	4 to 12
Assumed parameters		
Isentropic efficiency of gas turbine	η_{gt}	0.85
Isentropic efficiency of compressor	η_c	0.85
Combustion chamber efficiency	η_{cc}	1.0
Heat exchanger effectiveness	ε	0.9
Isentropic efficiency of steam turbine	η_{st}	0.95
Pump efficiency	η_{pump}	0.95
Volume ratio of diesel engine	r	20
Cut-off to stroke ratio of diesel engine	k	5%
Compressor inlet & ambient temperature	T_1 (*K*)	300
Mass flow rate of air to GTC	m_a (kg/s)	1
Mass flow rate of air to diesel engine	m_a (kg/s)	1
Lower calorific value of fuel to GTC	LCV (kJ/kg)	42000
Lower calorific value of Diesel	LCV (kJ/kg)	45500

3. Solution Technique

Using first and second laws of thermodynamics, parametric analyses of work output, thermal efficiency, SFC, and exhaust gases exergy losses of simple GTC, and the combined cycle has been carried out in the present study with certain assumptions listed in table 1. The pressure ratio and turbine inlet temperature of GTC are the variables to analyse and compare the proposed system with simple GTC and diesel engine in the range of 4 to 12 and 900K to 1200K respectively. Engineering Equation Solver (EES) is used to solve equations listed in section 4 based on assumptions and variables listed in table 1.

4. Mathematical modelling of Proposed system

For mathematical modeling of the proposed system shown in figure 1, the complete system is divide into three major sections. The major governing equations for the parameter under an analysis of the first one (diesel cycle), the second section (gas turbine cycle), and the third section (organic Rankine cycle) is listed below

Analysis of Diesel cycle

Heat supplied to Diesel cycle is given by $\left(\dot{Q}_s\right) = \dot{m}_g \, c_{pg} . (T_3 - T_2)$ (1)

Heat rejected by Diesel cycle is given by $\left(\dot{Q}_R\right) = \dot{m}_g \, c_{vg} . (T_4 - T_1)$ (2)

Net output from Diesel cycle $(W)_{netdiesel} = \dot{m}_g \, c_{vg} . \{\gamma_g (T_3 - T_2) - (T_4 - T_1)\}$ (3)

After putting the value T_2, T_3, T_4 in term of r, ρ and T_1 in equation (3) we get

$$(W)_{net} = T_1 . \dot{m}_g \, c_{vg} . \left\{\gamma_g . r^{(\gamma_g - 1)} (\rho - 1) - \left(\rho^{\gamma_g} - 1\right)\right\} \quad (4)$$

Where $\rho = 1 + k\,(r - 1)$ (5)

From equation (4) & (5)

$$(W_{net})_t = T_1 . \dot{m}_g \, c_{vg} . \left\{k . \gamma_g . r^{(\gamma_g - 1)} (r - 1) + 1 - \left[\left(1 + k\,(r - 1)\right)^{\gamma_g} - 1\right]\right\} \quad (6)$$

Temperature of exhaust gases from the Diesel cycle is

$$(T_4) = T_1 . \{1 + k\,(r - 1)\}^{\gamma_g} \quad (7)$$

Mass flow rate of flue in Diesel Engine is given by $(\dot{m}_f)_{diesel} = \dfrac{T_1 . (\dot{m}_a)_t \, c_{pg} . k . (r - 1) . r^{(\gamma_g - 1)}}{cv - c_{pg} T_1 . [1 + k\,(r - 1)] . r^{(\gamma_g - 1)}}$ (8)

Air Standard Thermal Efficiency of Diesel cycle $(\eta_{th})_{diesel} = 1 - \dfrac{\left[\left(1 + k\,(r - 1)\right)^{\gamma_g} - 1\right]}{k . \gamma_g . r^{(\gamma_g - 1)} (r - 1)}$ (9)

Analysis of Gas Turbine Cycle

Compressor Work is $(W_c) = \dot{m}_a . (h_7 - h_6)$ (10)

where $T_7 = T_6 \left\{1 + \left(r_p^{\alpha} - 1\right)/\eta_c\right\}$ and $\alpha = (\gamma_a - 1)/\gamma_a$ (11)

Work from gas turbine is $(W_{GT}) = \dot{m}_g . (h_9 - h_{10})$ (12)

where $T_{10} = T_9 \left\{1 - \eta_t \left(1 - r_p^{-\beta}\right)\right\}$ and $\beta = \left(\gamma_g - 1\right)/\gamma_g$ (13)

Mass flow rate of gasses is $\dot{m}_g = \dot{m}_a + \dot{m}_f$ (14)

where $\dot{m}_f = \dot{m}_a . \left\{\left(c_{pg} T_9 - c_{pa} T_8\right)/\left(LCV - c_{pg} T_9\right)\right\}$ (15)

The work output from GTC is

$$(W_{net})_{GTC} = (W_{GT}) - (W_c) \quad (16)$$

The GTC thermal efficiency is $(\eta_{th})_t = (W_{net})_t / Q$ (17)

The SFC of GTC is $(SFC)_t = \left(3.6 \times 10^5\right)/\left(\left(\eta_{th}\right)_t . LCV\right)$ (18)

For first heat exchanger (H.E.)

$$\dot{m}_a . (h_8 - h_7) = \varepsilon_1 . \dot{m}_{gdiesel} . (h_4 - h_7) \quad (19)$$

Therefore

$$T_8 = (1-\varepsilon_1)T_7 + \varepsilon_1.(\dot{m}_{gdiesel}/\dot{m}_a).(c_{pg}/c_{pa})T_4 \quad (20)$$

$$T_5 = T_4 - (\dot{m}_a/\dot{m}_{gdiesel}).(c_{pa}/c_{pg}).(T_8 - T_7) \quad (21)$$

Analysis of Organic Rankine Cycle

Mass flow rate of refrigerant is $\dot{m}_{ref} = \varepsilon.m_g.(h_{10}-h_{11})/(h_{12}-h_{15})$ (22)

Work output from the turbine is $W_{ST} = \dot{m}_{ref}.\eta_{ST}.(h_{12}-h_{13})$ (23)

Work required to run the pump $W_{pump} = \dot{m}_{ref}\,\nu_f.(p_{12}-P_{14})$ (24)

Work output is $(W_{net})_{ORC} = (W_{ST}) - (W_{pump})$ (25)

Analysis of combined Cycle

The combined cycle work output is $(W_{net})_{comb} = (W)_{netdiesel} + (W_{net})_{GTC} + (W_{net})_{ORC}$ (26)

Combined cycle thermal efficiency is

$$(\eta_{th})_{comb} = \left[(W_{net})_{comb} / \left((\dot{m}_f)_{diesel} \times LCV_{diesel} + (\dot{m}_f)_{GTC} \times LCV_{GTC}\right)\right] \times 100 \quad (27)$$

combined cycle SFC is $(SFC)_{comb} = 3600 \times \left((\dot{m}_f)_{diesel} + (\dot{m}_f)_{GTC}\right) / (W_{net})_{comb}$ (28)

5. **Result and discussion**

The pressure ratio and turbine inlet temperature (TIT) is used to analyse the energetic and exergetic performance of the proposed system and comparative analysis of the proposed system with simple GTC and diesel engine power plant. In the proposed system the exhaust gases are used to heat the compressed air of the GTC cycle and the exhaust gases of the GTC cycle are used to heat the refrigerant in the ORC cycle. In this way, the heat gains by the compressed air of the GTC cycle from the exhaust gases of the diesel cycle helps to improve the GTC cycle efficiency by increasing the combustion chamber air inlet temperature. On the other hand, the heat grains by the refrigerant of the ORG cycle from the exhaust gasses of the GTC cycle helps to enhance the work output of combined GTC-ORC. The effect of pressure ratio (r_p) and turbine inlet temperature (*TIT*) on the thermal efficiency, work output, SFC, and exhaust gas exergy losses of the combined cycle are examined parametrically using the first and second law of thermodynamic.

The variation of work output of diesel cycle, simple GTC plus diesel cycle, and combined cycle (proposed system) with respect to a pressure ratio (r_p) and turbine inlet

temperature (*TIT*) of GTC cycle in figure 3 (a) and 3(b). It is clear from these figures that the work output of the combined cycle as well as simple GTC is significantly affected by *TIT* and r_p. It is also observed that the work output of the combined cycle as well as simple GTC decreases with r_p and increases with an increase in *TIT*. The work output of the diesel cycle is independent of *TIT* and r_p as clear from equation (3) and due to this, it is not affected by *TIT* and r_p. Figure 3 (a) and (b) also indicates that the work output of the diesel cycle as well as summation of work output from diesel and GTC cycle is much less as compared to the proposed system. The work output of the proposed system at r_p=4, is 35.6% and 69.6% more as compared to the work output of diesel cycle when *TIT* = 900K and *TIT* = 1200K respectively. At r_p=12, it is 17.6% and 60.9% when *TIT* = 900K and *TIT* = 1200K respectively more as compared to work output of diesel cycle.

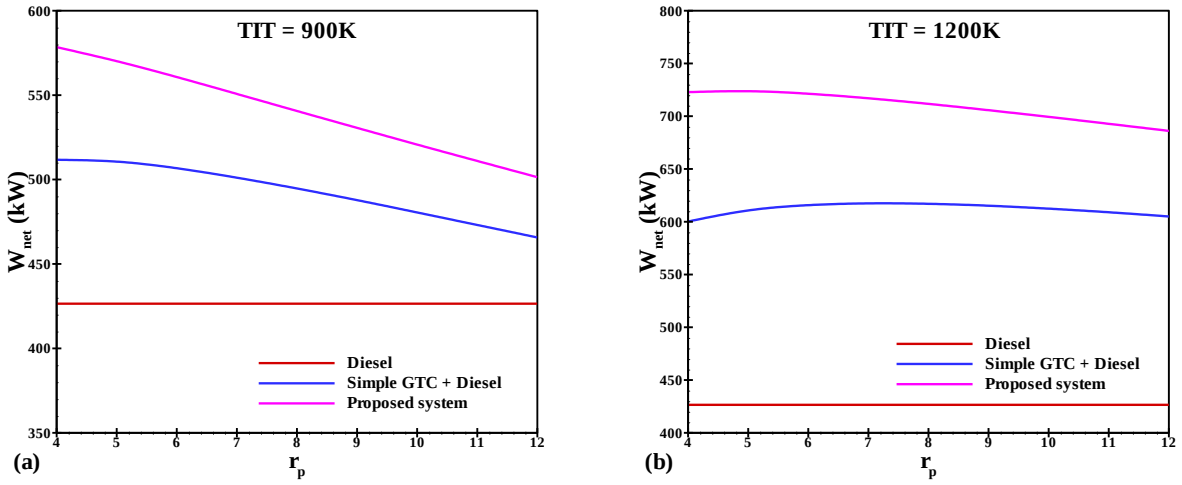


Figure 3. Variation of work output of diesel cycle, simple GTC plus diesel cycle and combined proposed system with respect to r_p at (a) *TIT* = 900K and (b) *TIT* = 1200K.

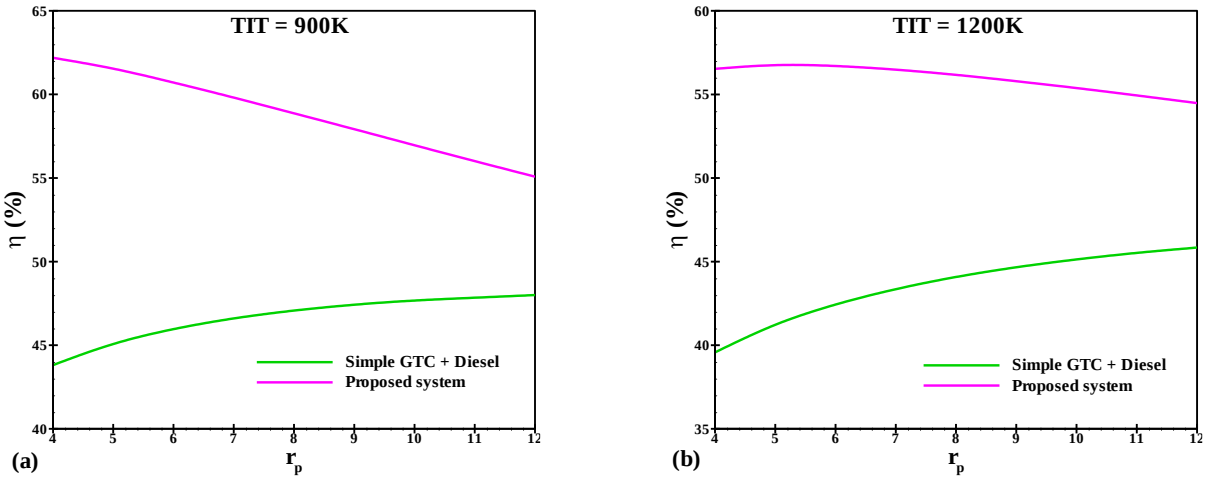


Figure 4. Variation of thermal efficiency of simple GTC plus diesel cycle and combined proposed system with respect to r_p at (a) TIT = 900K and (b) TIT = 1200K

From figure 3(a) and 3(b), at r_p=4, the work output of the proposed system is 13.1% and 20.5% more as compared to the work output of simple GTC plus diesel cycle when TIT = 900K and TIT = 1200K respectively. Whereas at r_p=12, it is 7.7% and 13.5% when TIT = 900K and TIT = 1200K respectively more as compared to work output of simple GTC plus diesel cycle. Figure 4(a) and 4(b) shows the variation of thermal efficiency of simple GTC plus diesel cycle and combined cycle (proposed system) with respect to a pressure ratio (r_p) and turbine inlet temperature (TIT) of GTC cycle. It noted from these figures that the thermal efficiency of the proposed system decreases with pressure ratio and increases with turbine inlet temperature (TIT) of the GTC cycle whereas, the thermal efficiency of simple GTC increases with pressure ratio and turbine inlet temperature (TIT) of GTC cycle. From figure 3(a) and 3(b), the work output of the diesel cycle is independent of the pressure ratio (r_p) and turbine inlet temperature (TIT) of the GTC cycle, this means that the thermal efficiency of the diesel cycle is also independent of pressure ratio (r_p) and turbine inlet temperature (TIT) of GTC cycle. From figure 4(a) and 4(b), at r_p=4, the thermal efficiency of the proposed system is 42% and 43% more as compared to work thermal efficiency of simple GTC plus diesel cycle when TIT = 900K and TIT = 1200K respectively. Whereas at r_p=12, it is 14.8% and 18.9% when TIT = 900K and TIT = 1200K respectively more as compared to the thermal efficiency of simple GTC plus diesel cycle.

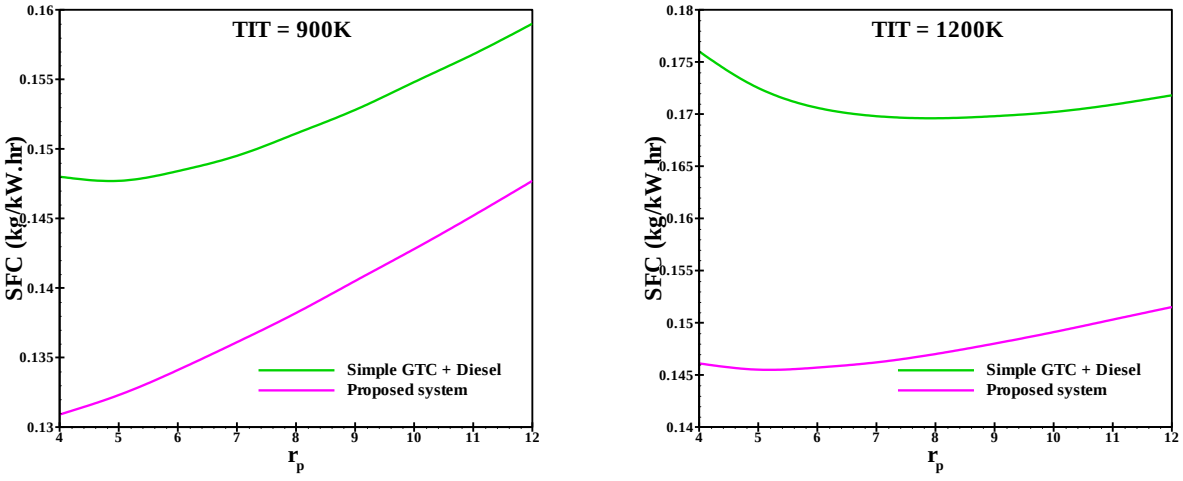


Figure 5. Variation of SFC of simple GTC plus diesel cycle and combined proposed system with respect to r_p at (a) TIT = 900K and (b) TIT = 1200K

The SFC of any cycle defines, as the ratio of the mass flow rate of fuel to work output of the cycle, this means that if the work output of the cycle increases, SFC decreases and vice versa. Figure 5 (a) and 5(b) shows the variation of SFC of simple GTC plus diesel cycle and combined

cycle (proposed system) with respect to r_p and *TIT* of GTC cycle. It observed that the SFC of the proposed system increase with pressure ratio because the work output of the proposed system decreases with r_p as shown in figure 3(a) and 3(b). Also, the SFC of the proposed system decreases with the turbine inlet temperature (*TIT*) of the GTC cycle. The SFC of simple GTC increases with pressure ratio and turbine inlet temperature (*TIT*) of the GTC cycle. From figure 5(a) and 5(b), at r_p=4, the SFC of the proposed system is 11.5% and 17% less as compared to work thermal efficiency of simple GTC plus diesel cycle when *TIT* = 900K and *TIT* = 1200K respectively. Whereas at r_p=12, it is 7.1% and 11.8% when *TIT* = 900K and *TIT* = 1200K respectively more as compared to the thermal efficiency of simple GTC plus diesel cycle.

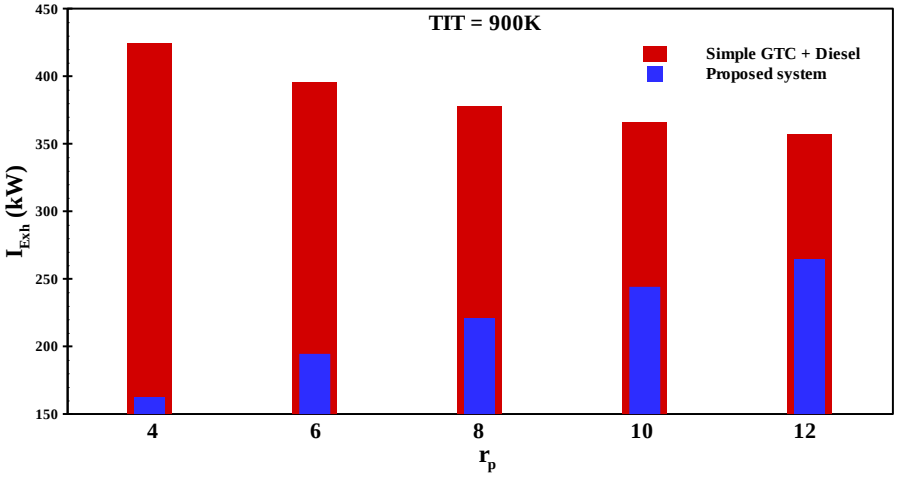


Figure 6 (a). Variation of exhaust gases exergy loss of simple GTC plus diesel cycle and combined proposed system with respect to r_p at *TIT* = 900K

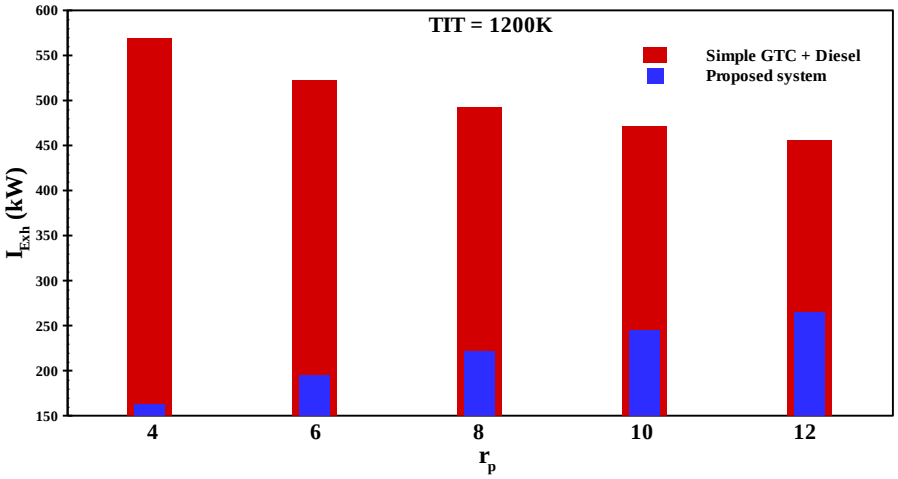


Figure 6 (b). Variation of exhaust gases exergy loss of simple GTC plus diesel cycle and combined proposed system with respect to r_p at *TIT* = 1200K

The exhaust gas exergy loss is significantly increased or decreased with an increase or decrease in the temperature of exhaust gases. Figure 6(a) and 6(b) shows the variation of exhaust gas exergy losses of simple GTC plus diesel cycle and combined cycle (proposed system) with respect to a pressure ratio (r_p) and *TIT* of GTC cycle. The exhaust gas exergy losses of the proposed system increase with pressure ratio because the temperature of exhaust gasses leaving to the environment from diesel engine increases with pressure ratio as the difference of T_8 and T_7 decreases with pressure ratio. Also, the exhaust gases exergy loss of the proposed system is not significantly affected by the turbine inlet temperature (*TIT*) of the GTC cycle. The SFC of simple GTC increases with pressure ratio and *TIT* of the GTC cycle. From figure 6(a), the exhaust gas exergy losses of the proposed system is 2.615 times at r_p=4 and 1.35 times at r_p=12 less as compared to simple GTC plus diesel cycle whereas in figure 6(b) the exhaust gases exergy loss of the proposed system is 3.5 times at r_p=4 and 1.72 times at r_p=12 less as compared to simple GTC plus diesel cycle.

6. **Conclusion**

In the present study, a thermodynamic analysis of the proposed system has been conducted. Parametric analyses have been done by varying the pressure ratio of GTC cycle from 4 to 14, and turbine inlet temperature 900*K* to 1200*K*. Based on the analyses, the following are the conclusions are finalized:

1. The work output of the proposed system plant increases by 13% to 20.4%, when TIT increases from 900K to 1200K.
2. The thermal efficiency of the proposed system plant increases by 15% to 43%, with an increase of TIT from 900K to 1200K.
3. The SFC of the proposed system is much less as compared to the sum of SFC of diesel and GTC system. The SFC of proposed system is 7.1% to 11.8% is less to the sum of SFC of diesel and GTC system.
4. The total exhaust gases exergy loss from the proposed system is decreased by 3.5 to 1.35 times as compared to the overall exhaust gases exergy loss from diesel cycle and simple GTC.
5. Overall the proposed system is much better in terms of energy as well as exergy performance as compared to simple diesel cycle and simple GTC cycle.

7. References

[1] A.M. Alklaibi, M.N. Khan, W.A. Khan, Thermodynamic analysis of gas turbine with air bottoming cycle, Energy. 107 (2016). https://doi.org/10.1016/j.energy.2016.04.055.

[2] M. Colera, Á. Soria, J. Ballester, A numerical scheme for the thermodynamic analysis of gas turbines, Appl. Therm. Eng. 147 (2019) 521–536. https://doi.org/10.1016/j.applthermaleng.2018.10.103.

[3] M. Maheshwari, O. Singh, Comparative evaluation of different combined cycle configurations having simple gas turbine, steam turbine and ammonia water turbine, Energy. (2019). https://doi.org/10.1016/j.energy.2018.12.008.

[4] V. Dolz, R. Novella, A. García, J. Sánchez, HD Diesel engine equipped with a bottoming Rankine cycle as a waste heat recovery system. Part 1: Study and analysis of the waste heat energy, Appl. Therm. Eng. (2012). https://doi.org/10.1016/j.applthermaleng.2011.10.025.

[5] R. Saidur, M. Rezaei, W.K. Muzammil, M.H. Hassan, S. Paria, M. Hasanuzzaman, Technologies to recover exhaust heat from internal combustion engines, Renew. Sustain. Energy Rev. (2012). https://doi.org/10.1016/j.rser.2012.05.018.

[6] M. Fallah, H. Siyahi, R.A. Ghiasi, S.M.S. Mahmoudi, M. Yari, M.A. Rosen, Comparison of different gas turbine cycles and advanced exergy analysis of the most effective, Energy. (2016). https://doi.org/10.1016/j.energy.2016.10.009.

[7] Sanjay, Investigation of effect of variation of cycle parameters on thermodynamic performance of gas-steam combined cycle, Energy. (2011). https://doi.org/10.1016/j.energy.2010.10.058.

[8] T. K. Ibrahim, M.M. Rahman, Effect of Compression Ratio on Performance of Combined Cycle Gas Turbine, Int. J. Energy Eng. (2012). https://doi.org/10.5923/j.ijee.20120201.02.

[9] M.N. Khan, I. Tlili, Performance enhancement of a combined cycle using heat exchanger bypass control: A thermodynamic investigation, J. Clean. Prod. 192 (2018). https://doi.org/10.1016/j.jclepro.2018.04.272.

[10] T.K. Ibrahim, F. Basrawi, O.I. Awad, A.N. Abdullah, G. Najafi, R. Mamat, F.Y. Hagos, Thermal performance of gas turbine power plant based on exergy analysis, Appl. Therm. Eng. 115 (2017) 977–985. https://doi.org/10.1016/j.applthermaleng.2017.01.032.

[11] M. Ghazikhani, I. Khazaee, E. Abdekhodaie, Exergy analysis of gas turbine with air bottoming cycle, (2014) 1–9.

[12] M. Ghazikhani, M. Passandideh-Fard, M. Mousavi, Two new high-performance cycles for gas turbine with air bottoming, Energy. 36 (2011) 294–304. https://doi.org/10.1016/j.energy.2010.10.040.

[13] O.K. Singh, Performance enhancement of combined cycle power plant using inlet air cooling by exhaust heat operated ammonia-water absorption refrigeration system, Appl. Energy. (2016). https://doi.org/10.1016/j.apenergy.2016.08.042.

[14] M.N. Khan, I. Tlili, W.A. Khan, Thermodynamic Optimization of New Combined Gas/Steam Power Cycles with HRSG and Heat Exchanger, Arab. J. Sci. Eng. 42 (2017). https://doi.org/10.1007/s13369-017-2549-4.

[15] H. Yağli, Y. Koç, A. Koç, A. Görgülü, A. Tandiroğlu, Parametric optimization and exergetic analysis comparison of subcritical and supercritical organic Rankine cycle (ORC) for biogas fuelled combined heat and power (CHP) engine exhaust gas waste heat, Energy. (2016). https://doi.org/10.1016/j.energy.2016.05.119.

[16] W. Zhang, L. Chen, F. Sun, Power and efficiency optimization for combined Brayton and inverse Brayton cycles, Appl. Therm. Eng. (2009). https://doi.org/10.1016/j.applthermaleng.2009.02.011.

[17] M.N. Khan, Energy and Exergy Analyses of Regenerative Gas Turbine Air-Bottoming Combined Cycle: Optimum Performance, Arab. J. Sci. Eng. (2020). https://doi.org/10.1007/s13369-020-04600-9.
[18] R.S.Mishra, M. Kumar, Thermodynamic models for combined cycle power plants used in organic Rankine and Brayton cycles, Int. J. of research in Engineering and Innovation.
[19] J. Galindo, J.R. Serrano, V. Dolz, P. Kleut, Brayton cycle for internal combustion engine exhaust gas waste heat recovery, Adv. Mech. Eng. (2015). https://doi.org/10.1177/1687814015590314.
[20] S. Zhu, K. Deng, S. Qu, Energy and exergy analyses of a bottoming rankine cycle for engine exhaust heat recovery, Energy. (2013). https://doi.org/10.1016/j.energy.2013.06.031.
[21] Y. Cao, Y. Gao, Y. Zheng, Y. Dai, Optimum design and thermodynamic analysis of a gas turbine and ORC combined cycle with recuperators, Energy Convers. Manag. (2016). https://doi.org/10.1016/j.enconman.2016.02.073.
[22] W. Sun, X. Yue, Y. Wang, Exergy efficiency analysis of ORC (Organic Rankine Cycle) and ORC-based combined cycles driven by low-temperature waste heat, Energy Convers. Manag. (2017). https://doi.org/10.1016/j.enconman.2016.12.042.
[23] A. Sadreddini, M. Fani, M. Ashjari Aghdam, A. Mohammadi, Exergy analysis and optimization of a CCHP system composed of compressed air energy storage system and ORC cycle, Energy Convers. Manag. (2018). https://doi.org/10.1016/j.enconman.2017.11.055.
[24] T. k. Ibrahim, M. Kamil, O.I. Awad, M.M. Rahman, G. Najafi, F. Basrawi, A.N. Abd Alla, R. Mamat, The optimum performance of the combined cycle power plant: A comprehensive review, Renew. Sustain. Energy Rev. (2017). https://doi.org/10.1016/j.rser.2017.05.060.
[25] H. Rostamzadeh, M. Ebadollahi, H. Ghaebi, M. Amidpour, R. Kheiri, Energy and exergy analysis of novel combined cooling and power (CCP) cycles, Appl. Therm. Eng. (2017). https://doi.org/10.1016/j.applthermaleng.2017.06.011.
[26] A.A.A. Abuelnuor, K.M. Saqr, S.A.A. Mohieldein, K.A. Dafallah, M.M. Abdullah, Y.A.M. Nogoud, Exergy analysis of Garri "2" 180 MW combined cycle power plant, Renew. Sustain. Energy Rev. (2017). https://doi.org/10.1016/j.rser.2017.05.077.
[27] M.N. Khan, Energy and Exergy Analyses of Regenerative Gas Turbine Air-Bottoming Combined Cycle: Optimum Performance, Arab. J. Sci. Eng. (2020). https://doi.org/10.1007/s13369-020-04600-9.
[28] K. Rahbar, S. Mahmoud, R.K. Al-Dadah, N. Moazami, S.A. Mirhadizadeh, Review of organic Rankine cycle for small-scale applications, Energy Convers. Manag. (2017). https://doi.org/10.1016/j.enconman.2016.12.023.

A Comprehensive Overview of Classical and New Perceivable Spatial and Temporal Artifacts in Compressed Video Streams

Dr. Muhammad Uzair,
Faculty of engineering, Islamic University of Madinah, KSA
uzair91@hotmail.com, muzair@iu.edu.sa

Abstract: Modern video encoding tools use lossy coding schemes due to the constraints of power and bandwidth. These coding approaches also create spatial and temporal artifacts which ultimately decrease the received video quality. Understanding these classical and new artifacts, especially produced due to new video encoding tools are very important for the development of efficient video codec's which can avoid these artifacts on the encoder side, rather than compensating them on decoder side. There is still a lack of knowledge about new artifacts due to the new video coding tools, besides the classical artifacts, i.e., blocking, ringing, etc. Many existing papers in the literature describe these artifacts, but none of the paper discusses these artifacts comprehensively. Therefore, this research provides a comprehensive overview of all classical and new artifacts (spatial and temporal) produced by current and new video coding tools, i.e., SVC (scalable video coding) and MVC (multi view coding) by H.264, and NVC (next generation video codec's - H.265/HEVC). The paper also describes the effect of new coding tools on classical artifacts and how these artifacts are related to each other. Overall, this research will assist in the design of more efficient adaptive quantization algorithms and coding mechanisms to improve the video codec performance.

Keywords: Spatial, Temporal, Compressed, Video, Codec's, Artifacts, Blockiness

نظرة عامة شاملة على القطع الأثرية المكانية والزمانية الكلاسيكية والجديدة التي يمكن إدراكها في تدفقات الفيديو المضغوطة

الملخص: تستخدم أدوات تشفير الفيديو الحديثة مخططات تشفير ضياع بسبب قيود القدرة وعرض النطاق. تؤدي أساليب الترميز هذه أيضًا إلى إنشاء نتائج مكانية وزمانية تؤدي في النهاية إلى تقليل جودة الفيديو المستلم. يعد فهم هذه القطع الأثرية الكلاسيكية والجديدة، وخاصة التي يتم إنتاجها بسبب أدوات ترميز الفيديو الجديدة، أمرًا مهمًا للغاية لتطوير برامج ترميز فيديو فعالة يمكنها تجنب هذه القطع الأثرية على جانب التشفير، بدلاً من تعويضها من جانب وحدة فك التشفير. لا يزال هناك نقص في المعرفة حول القطع الأثرية الجديدة بسبب أدوات ترميز الفيديو الجديدة، إلى جانب المصنوعات الكلاسيكية، مثل الحجب، والرنين، وما إلى ذلك. تصف العديد من الأوراق الموجودة في الأدبيات هذه القطع الأثرية، ولكن لم يناقش أي من هذه القطع الأثرية بشكل شامل. لذلك، يقدم هذا البحث نظرة عامة شاملة على جميع القطع الأثرية الكلاسيكية والجديدة (المكانية والزمانية) التي تنتجها أدوات ترميز الفيديو الحالية والجديدة، مثل SVC (ترميز الفيديو القابل للتطوير) وMVC (ترميز متعدد العروض) بواسطة H.264 وNVC (الجيل القادم من برامج ترميز الفيديو -H.265 / HEVC.) تصف الورقة أيضًا تأثير أدوات الترميز الجديدة على القطع الأثرية الكلاسيكية وكيف ترتبط هذه القطع الأثرية ببعضها البعض. بشكل عام، سيساعد هذا البحث في تصميم خوارزميات تكميم أكثر كفاءة وآليات تشفير لتحسين أداء ترميز الفيديو.

Introduction: Digital video data gets distorted during acquisition, compression, transmission, decoding and reproduction. The video quality is degraded during the quantization process (lossy compression) to meet the bandwidth, power, and time requirements. Quantization step sizes, especially large, could reduce power consumption, encoding time, and bandwidth requirements, but results in video quality degradation. The perception of the human visual system (HVS) plays an important role during lossy compression as different kinds of spatial and temporal distortions are related to the properties of the HVS. Higher compression ratios are achieved by reducing or eliminating that information which is not noticeable by HVS in order to catch up increasing demand for current industrial video communication system. The quality of the service should be monitored at the receiver side in order to maintain and improve the quality of the received video data, as the demand for a better quality has been higher than ever before. The poor video quality at the consumers end has also resulted in revenue lost in digital communication industry [1].

The compression techniques introduce many visual artifacts (spatial and temporal) in a compressed video. The spatial distortion is visualized in individual frames, while temporal distortion is observed when the video is played. Both spatial and temporal artifacts have many shapes and kinds of distortions [1]. The nature of the artifacts also depends upon which codec is used (MPEG x, H.26x), error position in a frame and whether error concealment is used at decoder or not, etc [2]. In addition to spatial and temporal artifacts, the artifacts produced during acquisition (e.g., camera motion blur), during transmission (e.g., freezing, packet loss), and during video post processing and display (e.g., spatial scaling, chromatic aberration) are also produced. However, these artifacts are not produced during compression [3, 4].

With new coding tools new spatial and temporal artifacts emerged which needs to be addressed, i.e., origin of the artifacts should be known in order to compensate them, etc. Many papers have been written in the past which discusses about spatial and temporal artifacts [3, 4, 5, 6, 7, 8, 9, 10]. However, these papers do not discuss the spatial and temporal artifacts comprehensively. Papers [3, 5, 6] discuss only a few of spatial and temporal artifacts, while paper [7] discuss only the compression artifacts. Similarly, paper [8] discuss only a few of temporal artifact and paper [4] describes only classical spatial and temporal artifacts. Moreover, most of the paper do not discuss new artifacts generated by SVC/MVC or NVC (H.265), except [7]. The existing literature does not discuss all the classical and new artifacts (spatial and temporal) comprehensively in a compressed video. However, this research provides a comprehensive overview and analysis of all spatial and temporal artifacts produced

by current and new video coding tools, i.e., MPEG-2, MPEG-4 Part 2, H.264, VC-1, SVC and MVC by H.264, and H.265. The paper also discusses the effect of new coding tools on classical artifacts and explains that how one artifact is related to other artifacts, i.e., masks/creates other artifact, etc. The paper also describes different ways that how these artifacts can be minimized, i.e., need of compensating artifacts at decoder side can be eliminated, etc.

The overall structure of the research is shown in Figure 1.

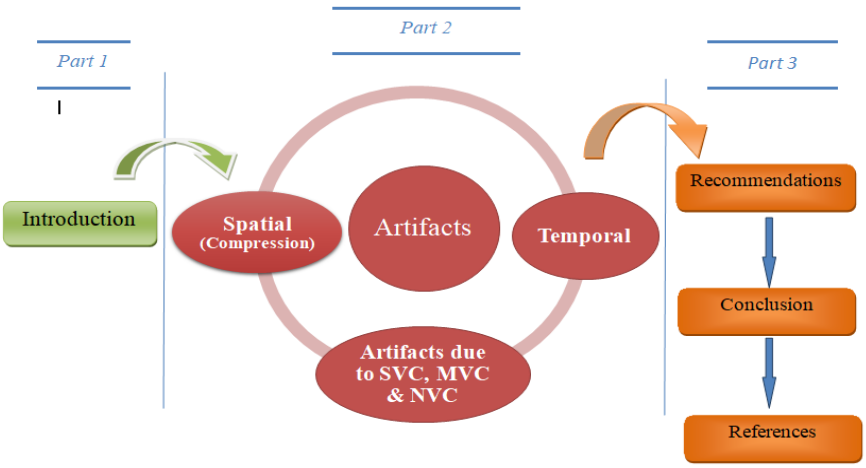


Figure 1: Overall structure of research

Following section presents the spatial artifacts.

1. Spatial (compression based) artifacts

The compression techniques employed in various existing and new video codec's are very similar, i.e., block based DCT (discrete cosine transformation) coding with motion compensation and subsequent quantization. In this kind of coding, quantization errors in the transform domain is the main reason of different kinds of artifacts. The encoding process is also affected indirectly through the quantization scale factor because of motion prediction and decoder buffer size [1]. Different kinds of spatial artifacts are discussed below.

1.1 Blocking artifacts: The blocking artifacts are the most prominent and annoying distortion in block based coding. Blockiness happens because of the two main reasons: 1) independent quantization of individual blocks which leads to discontinuities at the boundaries of adjacent blocks. 2) Due to the motion compensation (MC) prediction, as it reduces inter blocks correlation causing discontinuities near the blocks boundaries. Therefore, using the blocks of different sizes as basic units for transformation, quantization and motion compensation produce discontinuities across block boundaries [8, 9]. The amount and visibility of blocking artifacts increase at low bit rates, i.e., more compressed videos. Blocking artifact is more noticeable in the smoothly textured region, and generally hidden in more spatially active areas (especially

when coarse quantization is not used), or bright/dark areas. The lower order DCT coefficient (DC coefficient) is very important to determine whether a blocking distortion will be visible or not [4, 5]. Similarly, natural edges are very abrupt transitions while block edges are weak and regularly spaced. Blocking and slow motion may produce fine-granularity flickering [8].

1.1.1 Effects of macro blocking: For inter predictive coding; macro blocks are partitioned to 16x16, 16x8 and 8x16, down to 4x4 in H.264, H.265. This portioning allows to find better matches for each part by performing a separate motion search for each part [11]. If there is no de-blocking filter, these partitions may increase the blocking and also favor the appearance of motion compensation (MC) mismatch [7].

1.1.2 Effects of spatial scalability: For scalable video coding (SVC) in H.264, up sampling possibly increases blocking artifacts, as the block partitions of the lower layer are up sampled accordingly [7].

1.1.3 Blocking artifacts and H.265/HEVC: In H.264, the macro blocks had a maximum size of 16x16 pixels, and a limited number of sub-blocking patterns. In H.265/HEVC, the macro blocks can be 16x16, 32x32, or up to 64x64 luma sample. The coding efficiency also increases as the macro block size increases. The 64x64 luma size is also called coding tree unit (CTU) which is further divided into prediction units (PU) and transform units (TU) down to a size of 4x4. Due to the CTUs, H.265/HEVC provides new shapes and sizes including all the previous block sizes of H.264. The partitioning does not necessarily change the appearance of artifacts, as smallest size is equal to the H.264. However, blocking artifacts are still obvious, although smaller [2,7,12,13].

1.1.4 Blocking artifacts and different codec's: The blocking artifact appear in all block based video codec's, i.e., MPEG-2 Video, MPEG-4 Part 2, H.261, H.264 and VC-1, etc. MPEG-4 Part 2 and the MPEG-2 coded videos produce blockiness due to its transform size, i.e. an 8x8 DCT. However, the smallest transform size in H.264, H.265, and VC-1 is 4x4 which increases the blocking distortions as the number of block borders increase [7, 11]. Similarly, other block based coding techniques, i.e., vector quantization, block truncation coding, and fractal based compression also suffer from blocking artifact. However, blocking effects are reduced in JPEG 2000, i.e., wavelet transform based compression standards, as the transform is applied to the entire image instead of individual blocks [1].

1.1.5 Filtering blocking artifacts: For the minimization of blocking artifacts, de-blocking is done by the use of simple low pass filtering producing blurring even at those boundaries which do not have blockiness. MPEG-2 Video and MPEG-4 Part 2 perform blockiness compensation

at decoder side, as they do not have built in de-blocking filter, while H.264 and VC-1 use an in-loop de-blocking filter to minimize blocking [7, 8]. H.265 only applies the de-blocking filter to 8×8 sample instead of 4*4 as used in H.264, which reduces computational complexity and improves parallel processing operation [2, 12]. In H.264, the reduction of blocking may also produce flickering artifact [8].

1.2 Blurring: Blurring refers to a loss of spatial features and a decrease in edges or texture sharpness due to the removal of high frequency coefficients by quantization or edge-attenuating filters [3]. The coarse quantization further increases blurring. Similarly, de-blocking operators (low pass filters) used by different codec's, i.e., H.264/AVC and HEVC, etc., to reduce blockiness also produces perceptual blurring effect. A video with blurring artifacts due to de-blocking filters appears more pleasant to an observer than without filtering the video [6]. Blurring can also occur due to source related error. In H.264, the reduction of blocking and blurring may produce flickering artifact also [8]. Also, blurring in high spatial activity area may coincide with blocking and mosaic pattern effect as like DCT basis pattern [4].

1.2.1 Blurring and different codec's: The standards, i.e., MPEG-2, MPEG-4 Part 2, H.264 and VC-1 use course quantization which produce blur, and these standards have no integrated filter and use algorithms for de-blurring on the decoder side, if desired. Similarly, producing high frequency components at decoder side (inverse filtering) may yield over sharpening artifacts or introduce noise. The camera motion blur has similar effects like blur, but requires different approach for de-blurring [5, 7].

1.3 Ringing: This artifact is due to the removal of high frequency coefficients because of coarse quantization. It is perceived as ripples and overshoots near high contrast edges/lines and is more prominent in the wavelet coders. It is also known as the Gibbs phenomenon in one dimensional Fourier analysis. It is more visible near sharp and strong edges/lines and those areas where visual masking is weak, i.e., near smooth edges/lines. Both components, i.e., luminance and chrominance, suffer ringing artifacts. It is also not related to the blocking which depends on the existence of uniform/smooth areas, while ringing depends on the amount and strength of edges. A mosquito effect (temporal artifact) is observed by combining ringing with the motion of objects in successive frame [3, 13, 9]. The ringing is more annoying at low bit rate, but it is also observed in low compressed videos [7]. Figure 2 shows ringing artifacts [9].

Figure 2. Ringing artifact (a) Reference frame; (b) Compressed frame with ringing artifact [9]

1.3.1 Ringing and different codec's: Ringing is generally reduced with the use of in loop filtering in all video standards, i.e., MPEG-4 Part 2, VC-1, and H.264 and H.265, etc. As 4x4 transform size is used in H.264, VC-1, and H.265, ringing is reduced within one transformed block due to the limited space for over and undershooting [4, 14, 15]. However, the introduction of a 16x16, 32x32 and 64x64 macro block sizes in H.265 may increase ringing artifacts as compared to H.264 codec [2,7, 12].

1.4 Staircase effect: Coarse quantization truncates the higher order basis images to zero and the reconstruction of an image by using lower frequency basis images is either horizontally or vertically oriented, i.e., generally not tuned to the represent diagonal edges/features [5]. Stair case effect typically happens when horizontal and vertical basis functions are not able to accurately represent steep edges due to which horizontal or vertical basis functions becomes more significant. Therefore, a stair case structure is produced along diagonal lines/curve when it get mixed with the false horizontal and vertical edges at block boundaries as shown in Figure 3 (rectangular region) [9]. This artifact is also closely related to ringing and it is more noticeable when the size of a macro block becomes equal to stair case step size. This artifact is also related to blocking and mosaic pattern by demonstrating discontinuities at block boundaries [4, 7].

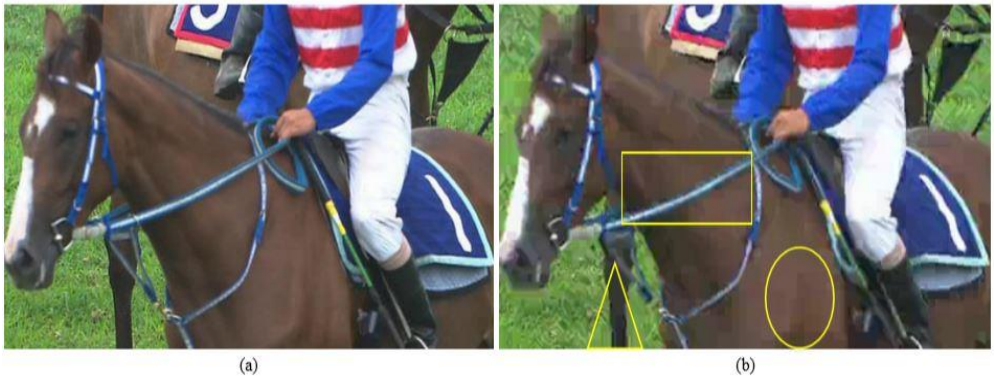


Figure 3. (a) Reference frame (b) Three different types of artifacts:
staircase effect (rectangular area); false edge (triangular area); mosaic pattern (elliptical area) [9]

1.5 False edges: It happens due to the transfer of blockiness distortion from reference to predicted frame due to motion compensation. It produces high frequency noise within the blocks as contrast to blocking which produces same noise along the block boundaries [3]. This effect is more noticeable in smooth areas and generally do not propagate to high spatial activity regions in predictive coded frames [4]. False edge appears near a true edge as shown in Figure 3 (triangular region) [9].

1.6 Mosaic effect (patterns): This effect usually happens when there is mismatch between adjacent blocks, i.e., luminance transitions. Generally, quantization makes all AC coefficients to zero within a block and each block is represented as a constant DC block. Therefore, DC value is different in each block, and by putting them together an abrupt luminance change creates a mosaic pattern. The mosaic pattern is highly visible at smooth regions, i.e., black/white boards, etc. Mosaic pattern is shown in Figure 3 (elliptical region) [9]. Mosaic pattern becomes visually more prominent when two adjacent blocks have different directional orientation. However, mosaic pattern generally reduces if there are many successive frames with high spatial activity area. This effect typically coincides with the blocking effect; however, blackness between two blocks does not always mean the presence of mosaic effect between same two blocks. In case of intra coded blocks, this effect has many similar kind of aspects as basis image effect. Moreover, it may also be introduced by the blocks suffering from the basis image effect [4, 6, 14].

1.7 Color bleeding: Color bleeding happens due to the removal of high frequency coefficients of the chroma components which leads to false color edges. After compression, smearing happens between color channels areas of strongly differing chrominance, as distortions are inconsistent. The resolution of the color channels (Cb and Cr) is half than luminance channel Y and due to the lower resolution interpolation is involved which further enhances inconsistent color spreading resulting in color bleeding [9]. Color bleeding can also happen even without chroma sub sampling. It may also occur due to the incoherent image rendering across the luminance and chrominance channels and can extend over an entire block due to the chroma sub sampling [7, 16]. Figure 4 [10] shows an example of color bleeding (rectangular region) showing chromatic distortion and inconsistent color spreading.

Figure 4. Color bleeding artifact (a) Reference frame (b) Compressed frame with color bleeding [10]

In chroma sub sampled images, one blurred chroma sample may also extend across multiple luma samples and in such case blurring is also called as color bleeding. Although, chroma sub sampling increases the perceived strength of the color bleeding; color bleeding can also propagate across views, especially using checkerboards arrangements (MVC) [7].

1.8 DCT basis image effect: It happens due to the course quantization when a single DCT coefficient becomes dominant in a block with the reduction of all other coefficients [3]. This artifact has same origin as ringing, but it can also occur other than sharp edges or lines. This effect is also similar to stair case artifacts. Figure 5 shows a basis pattern effect (rectangular regions) [9]. The visibility of the DCT basis pattern also depends on the nature of the texture region (where usually it occurs) [4, 17].

Figure 5: Basis image effect. (a) Reference frame (b) Compressed frame with basis pattern effect [9]

Due to this effect, when a smooth block is coded it might exhibit strong blocking and blurring artifacts along mosaic pattern effect in adjacent blocks. Moreover, just like mosaic pattern, this effect also decreases over time in high spatial activity area due to the accumulation and refinement of higher order AC coefficients [4, 8].

1.9 False contouring: False contouring can occur as a result of insufficient quantization of DC and low order AC coefficients or their inappropriate distribution. In smoothly texture areas, the pixels change their values gradually and by directly quantizing pixel values in these areas false contouring appears like a step like gradations and may effect to the whole block [4].

1.10 Clipping: It is the truncation of the image values (luminance and chrominance) during arithmetic precision producing abrupt cutting of the peak values at the top and bottom, i.e., creating an aliasing artifacts at peaks as caused by the high frequencies. Peaking, i.e., a sharpness enhancement technique can also produce clipping by adding positive and negative overshoots to the edges. For 8 bit precision, clipping is generally represented as 0 or 255, i.e., percentage of pixels having boundary values [16, 18, 19].

1.11 **Contrast**: It refers as the difference in the luminance value of a pixel of interest and the background. It highly depends on the ability to distinguish an object from its background which is called as dynamic range of a signal. The perception of contrast also varies from human to human as it also depends upon everyone's mind reference image about objects and sometimes about colors [16, 18, 19].

1.12 Sharpness: It refers to the clarity of details and contours of an image. It can be evaluated using the information provided by edges in the spatial domain or by using high frequencies in the transformed domain. It also highly depends on content, spatial resolution, contrast, and noise [16].

1.13 Noise: It is produced during random processes linked to transmission and generation techniques and is more visible in smooth regions or regions having small variations in the spatial or temporal dimension. Due to this effect, the details and quality of the image degrades.

1.14 Banding effect: Banding effect occurs in large and smooth regions in the reconstructed images when the large quantization step sizes are used. Figure 6 shows the banding effect [20].

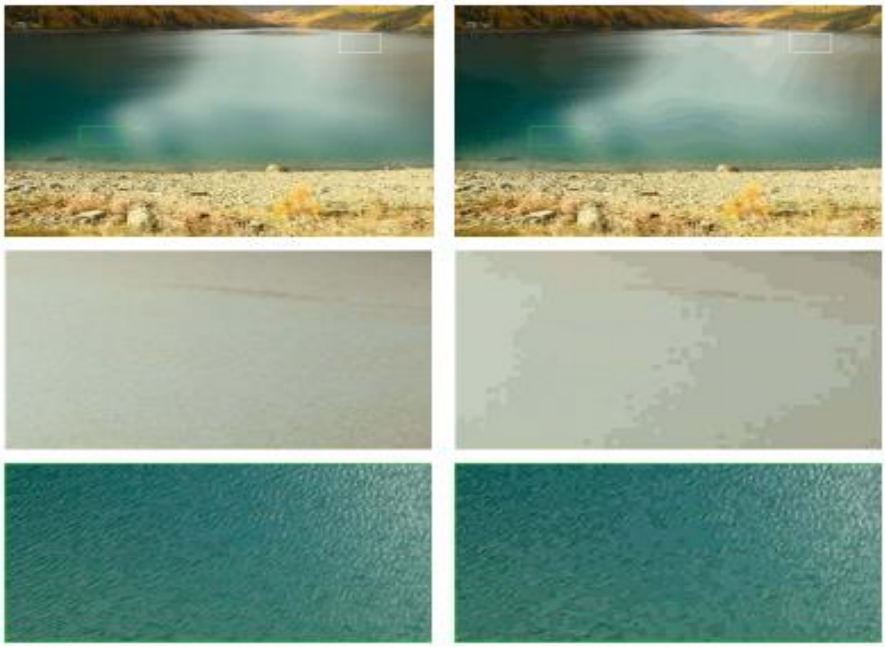


Figure 6: Banding effect (a) Reference frame (b) Restored image with banding effect [20]

The above section has comprehensively discussed the spatial artifacts describing their occurrence reason, relation with other artifacts, and their visual impacts. Figure 7 shows visual a visual pattern of different compression artifacts further highlighting the differences between spatial artifacts.

Figure 7: Different compression artifacts: blocking (marked as 1), blurring (marked as 2), ringing (marked as 3), stair cases (marked as 4), basis patterns (marked as 5), mosaïking (marked as 6) [7]

Table 1 shows a summary of the spatial artifacts [1, 4, 7, 9].

Table 1: Summary of spatial (compressed) artifacts

Artifact	Occurrence reason & Spatial extent within an image	Coexisting artifacts	Relation with other artifacts	Complete removal of a DCT coefficient (truncation effect) or effect of quantization process (association and continuity between blocks)
Blocking	Due to the independent quantization of adjacent blocks; 4*4 blocks up to 64*64 macro block (HEVC)		**1**. Without de-blocking filter, macro block partitioning increases the probability of blocking, and also favors the increase of MC mismatch. **2**. De-blocking filter produces blur. **3**. Blocking & slow motion may produce fine-granularity flickering. **4.** Reduction of the blocking in H.264 may give rise to flickering.	Effect of quantization process
Blurring	Loss of high frequency components; 4*4 block (H.264, H.265)	Ringing at sharp edges; color bleeding (chroma)	**1.** Reduction of the blurring in H.264 may give rise to flickering. **2**. Producing high frequency components at decoder side yield over sharpening artifacts or noise. **3.** Blurring in high spatial activity areas may coincide with blocking and mosaic pattern effect	Truncation effect
Ringing	Insufficient approximation of steep edges; 4*4 block (H.264, H.265)	Blurring	**1.** Mosquito effect is also related to the high frequency distortions as introduced by the ringing effect.	~

Staircase effect	Insufficient approximation of diagonal edge and other features; global spatial extent	Basis image effect for low quantization step sizes	**1**. Relates to blocking and mosaic pattern in terms of highlighting discontinuities between adjacent blocks. **2.** Closely related to ringing and becomes significant when stair case step size equals the size of a macro block through the influence of blocking	Effect of quantization process
False edges	Due to the transfer of blockiness distortion from reference to predicted frame due to MC		**1.** Coincides with the blocking distortion	~
Mosaic Patterns	Apparent mismatch between adjacent blocks due to the truncation of AC coefficients to zero; global spatial extent		**1.** May be produced by the basis image effect **2.** Coincides with blocking effect	Effect of quantization process
Color bleeding	Caused by the coarse quantization of high frequency components of chroma components; spatial extent: 64*64 macro block (HEVC)		**1**.In chroma sub sampled images, blurring is also called as color bleeding	Truncation effect
Basis Image effect	Loss of all but one DCT coefficients; spatial extent: 4*4 block (H.264, H.265)	Stair cases	**1**. May introduce mosaic patterns. **2**. May also introduce blocking and blurring along the boundaries in low spatial activity regions.	Truncation effect
False Contouring	Insufficient quantization of DC and low order AC coefficients or their inappropriate distribution; global spatial extent			Effect of quantization process

*hyphens tell that the distortion does not depend on any other artifact directly or indirectly.

The next section describes the temporal artifacts.

2. Temporal Artifacts

Temporal artifacts are mainly produced when the same region of a frame is coded inconsistently in successive frames of a video sequence. This can happen due to the changes in the type of predictions, quantization levels, motion compensation, or combination of these factors [21, 22]. Temporal artifacts are observed during video playback. It is important to know that many new artifacts have emerged with the advent of new video codec's as contrast to the spatial artifacts. For an example, texture floating is more significant in H.264/AVC coded video as compared to the early standards. However, it is again minimized comprehensively in the H.265. Similarly, evaluating (objectively) temporal artifact is more difficult as compared to the spatial artifacts. Different kinds of temporal artifacts are described below.

2.1 Flickering: Flickering is a temporal artifact which has significant visual impact. It is basically the change in the luminance or chrominance values along temporal dimension, i.e., coarse quantization from frame to frame. It can also occur due to the variation of prediction techniques from frame to frame, i.e., different intra prediction modes between successive frames. It is mainly observed in static regions where it is more prominent in the background rather than those regions which are in motion. Flickering is generally observed in a low to medium rate coding. Although, blocking and blurring are reduced successfully in H.264, but may increase flickering in intra coded frames [8, 9, 23]. Sometimes, it is also difficult to describe this artifact because of its complexity and appearance variations. Due to the variations, flickering can be further divided into three types based on the appearance and locations, i.e., mosquito noise, fine-granularity and coarse-granularity flickering. Figure 8 shows an example of flickering artifacts [10].

Figure 8. Flickering artifact: Reference frame (left), compressed frame with flickering (right) [10]

2.1.1 Coarse granularity flickering: It refers to the sudden luminance variations in low frequency across Group of Pictures (GOPs. For I-P GOP structure, I frame is intra coded, while all P frames are inter coded using predictions. Between two consecutive GOP's, there is no relation between I frame in the current GOP and the last P frame in the previous GOP. Therefore, a sudden luminance change happens between these two frames (I and P), especially if the scene remains the same. The frequency of this artifact depends on the size of GOP. Using variable GOP lengths and employing a new I frame only when scene change occurs this artifact may be avoided or significantly reduced [1, 8, 9].

2.1.2 Fine granularity flickering: This artifact occurs due to slow motion and blocking distortion in large low to mid energy areas. As considerable blockiness occurs at each frame in low to mid energy regions and the blockiness and DC values also change frame by frame in the corresponding blocks due to the motion or texture details. This causes flashing of these regions at high frequencies, which is eye catching and perceptually annoying [9, 23].

2.1.3 Mosquito noise: Mosquito noise is around sharp object boundaries and is due to both ringing distortions and motion compensation mismatch error. It happens because same region of a frame is coded differently in successive frames. Its appearance is like mosquitoes flying around, i.e., moving along with moving objects, as the plane region has weak visual masking effect and moving objects attract more visual attention. Therefore, this effect is easily noticeable and has huge impact on the video quality. In terms of visibility, ringing distortion has less effect as compared to the motion estimation error. A mosquito like noise can also occur due to the encoder/decoder drift due to the finite precision of the floating point operations in encoding/decoding process. This kind of noise may also be visible at low compressed videos [4, 8, 9, 14, 17].

2.2 Jerkiness: The Jerkiness happens when the motion of the object appears discontinuous. This is due to the reason that speed of the moving object is higher than the available temporal resolution. Jerkiness may become more visible with the motion of strong objects in a frame. Traditionally, jerkiness thought due to the low temporal resolution of video acquisition device, or when some frames are dropped or delayed due to low bandwidth constraint and the video does not remain smooth. In new video coding standards, as frames are divided into layers with coarse to fine temporal resolutions. Due to bandwidth constraints if fine resolution layers need to be dropped, jerkiness might appear. It may also happen due to the transmission delays of the bit stream, i.e., decoder's ability to buffer against fluctuations. Jerkiness can be evaluated using frame rate and temporal activity [4, 5, 9].

2.3 Floating: It refers to the emergence of an illusive motion in certain areas different from their surrounding environment. This illusive movement is wrong as these areas should not move or move together with the background, but they create an illusion due to their movement and look like floating on the top of surrounding background. It happens due to the use of skip mode in video coding, i.e., copying a block in successive frames without updating their details. Figure 9 shows floating artifact [10]. Floating effect can be divided into texture and edge floating depending upon where it appears [9, 10].

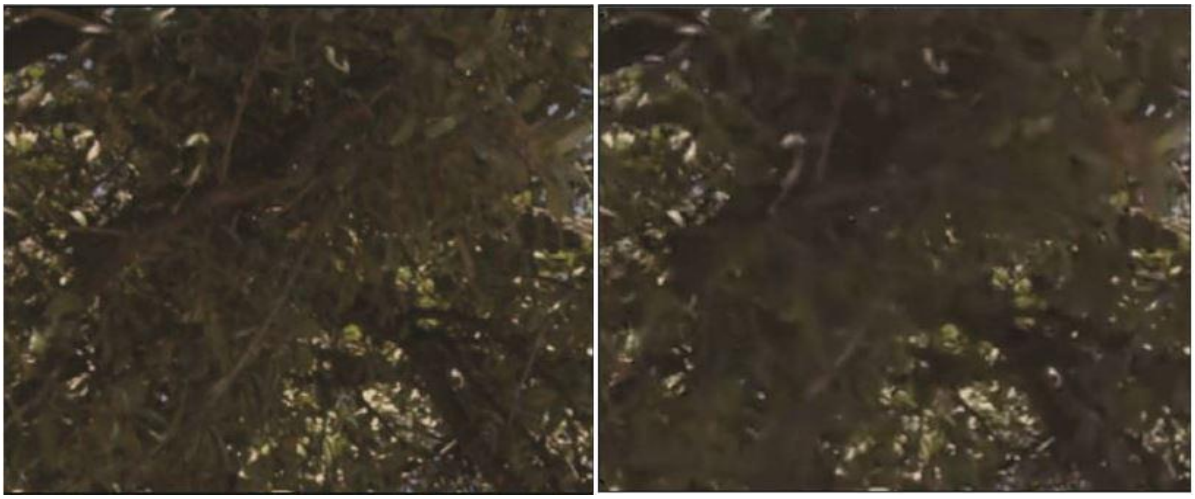


Figure 9. Floating artifact: Reference frame (left), Floating artifact (right) [10]

2.3.1 Ghosting (t**exture floating):** Ghosting generally happens when a scene having large mid energy textured area (trees or water surface) is captured using slow motion camera, i.e., relative motion between the floating regions and the background creates this effect. It appears like object persistence due to the planned temporal low pass filtering. Rather than actual shifting of image content, video encoder use skip mode for texture regions with zero motion to copy a block from one frame to another (traces of video content remains same in successive frames) in order to save bandwidth and an increase in the mean absolute error. However, this process creates a strong texture floating illusion in an opposite direction at same speed with respect to the camera motion. This effect is typically observed in high energy texture and edge regions. As textures are less visible in very bright or dark regions, therefore, the visibility of this effect is also limited by the luminance levels around the floating areas. This effect is also called texture floating in the literature [4, 5, 9].

2.3.2 Stationary area temporal fluctuations (edge neighborhood floating): This effect most likely also happens due to the use of skip mode and appears at stationary regions next to the boundaries of moving objects. In this case, stationary areas may also appear like a wrapped package surrounding and moving together with the object boundaries rather than remaining stationary. However, this effect may appear without global motion as contrast to the ghosting (texture floating). It is similar kind of fluctuations as associated with the mosquito effect in stationary areas containing major spatial activity, where it is difficult to notice minor differences between a region in one frame and the same region in the next frame of the sequence. This effect is also called as edge neighborhood floating in the literature [4, 7].

2.4 Motion compensation (MC) mismatch: It happens due to the inaccurate motion estimation (ME) during block based matching approach. This results in a mismatched spatial prediction with respect to the current macro block. Motion compensation is generally performed on the luminance value, where the motion vector indicates the spatial displacement of the current macro block from its prediction. Translational block based motion models and motion vectors can cause the inaccurate motion estimation and may produce a reconstructed video with highly visible distortions. In the worst case scenario when the prediction is truncated due to the high error, the reconstruction of the frame will happen with high non-correlation with the current frame. MC mismatch also produces mosquito noise. [24, 25].

2.5 Chrominance mismatch: Chrominance mismatch happens when the same motion vector is used for prediction which has been obtained by using luminance information. In this case, a macro block differs from its own general color and the color of the surrounding area. The chrominance mismatch generally does not appear at object boundaries; therefore, luminance prediction is satisfactory in these areas [3, 25].

2.6 Scene changes: Whenever there is a scene change, a sudden change in spatial features happens in frames before and after the scene. The perceived quality may degrade up to ten percent of the normal resolution after scene cut, although, may restore within half a second. As intra coding is used for the first frame for a new scene, predictive coding is not very efficient in such scenario. The quality of the initial frames is generally poor and builds up gradually as finer spatial characteristics predicatively accumulate. Therefore, there is always a loss of video quality whenever there is a scene change and is generally masked. This degradation in the quality becomes more noticeable when the video is displayed at low rates or as an individual images. One of the solutions is to mix the scenes before and after the cut by preprocessing the source by alpha-mixing. This step provides correlation among frames across the scene change

and also avoids the intra frame coding of the first frame after scene change and any kind of related distortions [1, 4].

2.7 Smearing: While recording a video sequence, the light from several moving objects is incorporated at a single point in the recording. Smearing occurs when the recorder simply cannot change the intensity of the beam fast enough to cope with the resolution. Smearing causes loss of spatial resolution and blurring of details. Smearing is visually more noticeable if the viewer is tracking the moving object while watching. The visibility of this effect also depends the speed of the moving object and other spatial masking effects [4, 6].

2.8 Down/up sampling: The up and down sampling of the spatial resolution is not directly related to the block based DCT approach. However, this process is used when the resolution of the source or the display device is not the optimum for coding the algorithm. The down sampling process discards the even fields and makes the vertical resolutions as half. There are many temporal distortions associated with the loss of even field, i.e., changes in vertical sizes, jitter and spatial variations. These artifacts would be generally masked by other spatial and temporal artifacts as they are only visible in small objects or with fine details [4].

2.9 Temporal pumping artifact: H.265/HEVC may use three kinds of prediction, i.e., intra only, low delay (LD) and random access (RA). Temporal flickering artifacts are easily noticeable for intra prediction approach. For LD and RA, the coding efficiency is related to the quantization parameter (QP), which may produce severe quality variations among adjacent pictures in a GOP leading to a perceptually temporal pumping effect (TPA) at medium and low bitrates [26, 27].

2.10 Jitter: This effect relates to the difference in end to end delay between selected packets in a sequence without taking into account any lost packets. This effect can also occur due to the absence of transmission errors creating a similar effect as a lost packet at decoder when the playoff time is missed [3, 16].

2.11 Frame freezing: Frame freezing happens when a single frame is repeated again and again on the screen permanently or often. Frame freezing can also be done on demand to enhance or highlight some thing. Frame freezing deteriorates the received video quality [28].

2.12 Frame skipping: Frame skipping happens when certain frames are not displayed in the video. Sometimes, it is done to improve certain performance, but visual smoothness deteriorates. Frame skipping can even happen due to the hardware failure [29].

2.13 Shimmering Video is encoded in a group of pictures (GOP) in which I frames are intra coded and have high quality, while P and B frames are intermediate frames. Due to the motion

estimation error, there can be a sudden drift (difference) between these frames which may increase the P and B frames between successive GOP's. This effect is called the shimmering [16].

2.14 Jagged motion: This effect also happens because of the inaccurate motion estimation. The motion estimation performs well when the movement of all pixels in a macro block is same. However, when the residual error due to motion estimation is large, it is coarsely quantized resulting in jagged motion [3].

2.15 Spread: It is due to the temporal propagation of error between frames due to the motion estimation error. If there is an error in a frame at the start of GOP, then the error spreads due to the prediction from the corrupted frame. It can drift up to whole GOP (until the new I frame in the next GOP), and causes a severe degradation in the reconstructed video.

2.16 Aliasing: When the spatial or temporal contents of a scene are more than the Nyquist rate, aliasing occurs. It can happen between two frames from different scenes, especially when there is a packet loss during scene change. However, the viewer may not notice it when the local movement is slow in the new scene. Temporal aliasing produces the wagon wheel effect, i.e., a situation where an apparent frequency of rotation has been changed by aliasing and a spoked wheel appears to rotate too slowly or even backwards [16, 29].

The above section has comprehensively discussed the temporal artifacts describing their occurrence reason, relation with other artifacts, and their visual impacts. Table 2 shows a summary of the temporal artifacts [1, 4, 7, 9].

Table 2: Summary of temporal artifacts

Artifact	Occurrence reason	Impact	Relation with other artifacts
Flickering	Coarse quantization from frame to frame	Significant impact on perceived video quality	Blocking and blurring are reduced in H.264, but may give rise to flickering
Coarse granularity flickering	Luminance fluctuations in low frequency across Group of Pictures (GOPs).	Frequency and impact depends on the size of GOP	
Fine granularity flickering	Due to slow motion and blocking distortion in low to mid energy regions	Highly noticeable and perceptually annoying	Blocking & slow motion may produce fine-granularity flickering
Mosquito noise	Quantization of high frequency components and MC errors	Strong impact on perceived quality	May also be produced by MC mismatch.

Jerkiness	Speed of the moving object is higher than the available temporal resolution	Become more visible with the motion of strong objects in a frame	
Floating	Encoder uses skip mode to copy a block from one frame to another (encoding of MC predicted residue is skipped)	Appears as an illusive motion in certain areas different from their surrounding environment.	
Ghosting (Texture floating)	Encoder uses skip mode and texture regions with zero motion to copy a block from one frame	More noticeable in high energy texture and edge regions	
Stationary area temporal fluctuations (edge neighborhood floating)	Happens due to the use of skip mode. Occurs at stationary regions next to moving objects.	Stationary areas may also appear like a wrapped package surrounding and moving together with the object boundaries	Similar kind of fluctuations as associated with the mosquito effect in stationary areas
Motion compensation (MC)	Happens due to the inaccurate motion estimation (ME)	High impact on visual quality.	MC mismatch also produces mosquito noise.
Chrominace mismatch	Due to use of same motion vector (obtained using luminance components) for chroma components	A macro block differs from its own general color and the color of the surrounding area	
Scene changes	A sudden change in spatial features in frames before and after the scene, whenever there is a scene change	More noticeable when the video is displayed at low rates or as an individual images.	
Smearing	Occurs when the recorder cannot change the intensity of the beam fast enough to cope with the resolution	Smearing causes loss of spatial resolution and blurring of the details	
Down/up sampling	Happens by discarding even fields during down sampling and making vertical resolutions as half	Causes jittery movement and spatial fluctuations	
Temporal pumping artifact	Happens when quantization varies significantly between adjacent pictures	Severe quality variations among adjacent pictures in a GOP	

The next section presents new artifacts (spatial and temporal) generated due to the use of new coding tools. The section also describes the effect of new coding tools on classical artifacts.

3. New artifacts (spatial and temporal) in new coding tools

Most of the artifacts (spatial and temporal) discussed in section 1 and 2 are present when a video sequence is coded using MPEG-2, MPEG-4 Part 2, H.264, VC-1, H.265. However, the use of new coding tools, i.e., scalable video coding (SVC), multi view coding (MVC), and NVC (next generation video coding) may also introduce new artifacts as described.

3.1 Effect of new transforms and transform sizes in new video standards: MPEG-2 and MPEG-4 Part 2 video standards have an 8x8 DCT transform size. However, the smallest transform size for H.264, H.265 and VC-1 is 4x4. With 4*4 sizes, the blocking distortion may increase as the number of block borders increase. Within one transformed blocks, smaller transform sizes (4*4) also reduce ringing due to the limited space for over and undershooting. H.264 and H.265 can switch between the two integer transform (4x4 and 8x8) while using the High profile. Although, basis functions of the 4x4 DCT and the integer transform are same, but in reality they produce different transform coefficients and transform coefficient distributions for a number of input signals except blocks with DC component only. Integer transform produce less transform coefficients as compared to the DCT avoiding the loss of signal energy during the inverse transform process. There is no current research available to find out that what kind of new artifacts can be produced due to new integer transform used in H.264 and H.265 [2, 7, 11]. H.265/HEVC uses 4*4 transform size, but with DST (discrete sine transform) for only luma blocks, as it better fits with the statistical property. Moreover, 4×4 DST transform is not computationally complex and demanding as compared to the 4×4 DCT transform and also provides approximately 1% bit rate reduction in intra frame coding [1, 2, 7, 12].

3.2 Effect of macro block partitioning: The effect of block partitioning has already been described in blocking and ringing distortion section, i.e., 1.1.1 & 1.3.1, respectively.

3.3 Effect of larger size macro block in H.265: The coding efficiency increases as the macro block size increase. However, the introduction of a 16x16, 32x32 and 64x64 macro block sizes in H.265 enhances the occurrence of ringing artifacts as compared to H.264 codec, which is due to the increased number of coefficients and samples available for over and undershooting [2,7,12].

3.4 Effect of different coding modes: Other than the number and shape of macro block partitions, new artifacts are also produced depending upon which coding mode is used. The flickering or pumping artifact is produced when coding mode of a certain frame area changes in successive frames. This is due to the fact that residuals from prediction differ greatly which produces different coded residual after quantization and produce flickering. The pumping

artifacts can be avoided by using similar coding modes for a region in successive frames [2, 7, 11].

3.5 Multi view coding (MVC): In this type of coding, the coding of multiple views of a scene is performed. This is done in order to obtain a three dimensional representation of a scene or part of it. This kind of coding also induces or favors many artifacts. MVC has three different approaches as described below [30, 31].

3.5.1 Depth map quantization: This approach produces these artifacts.

a) Depth ringing (depth bleeding) artifacts: The encoding of a two dimensional image or texture is a type of multi view coding, where depth map indicates the distance of each pixel from the camera. Depth maps are compressed like textures and produce similar artifacts like other codec's [1]. MPEG-4 part 2 explicitly specifies the depth maps coding. The quantization of depth maps yields depth ringing distortions of the depth map [7, 31]. It is also called depth bleeding and it is most noticeable at steep edges of the depth map as shown in Figure 8 [7].

b) Card board or puppet theater effect: Depth estimation error and harsh quantization may also produce card board or puppet theater effect as shown in Figure 10 [7]. As shown in the figure, light colors have more depth as compared to dark colors. The appearance of this artifact is like a two dimensional layers instead of smooth depth transitions, i.e., layer like depth map. The existence of depth map & texture coding may produce a superposition among them and they may mask each other [7, 31].

3.5.2 Frame packing artifacts: Frame packing is a stereoscopic video coding, i.e., second approach of MVC, which is available in H.264 and H.265. In this approach, a single view is used for coding the left and right views using supplemental enhancement information (SEI) message. The interleave coding, i.e., a kind of frame packing approach, of the two views can cause crosstalk of artifacts. The mosquito noise, pumping and MC mismatch may also increase due to the interleaving of views. Side by side and top bottom frame packing can also highlight these artifacts at the borders between the two views. Similarly, column and row alternation as well as checkerboard arrangements can also cause crosstalk. In check board arrangements, color bleeding can also propagate across views. Similarly, using frame alternation provision, MC mismatches increases [7, 31, 32].

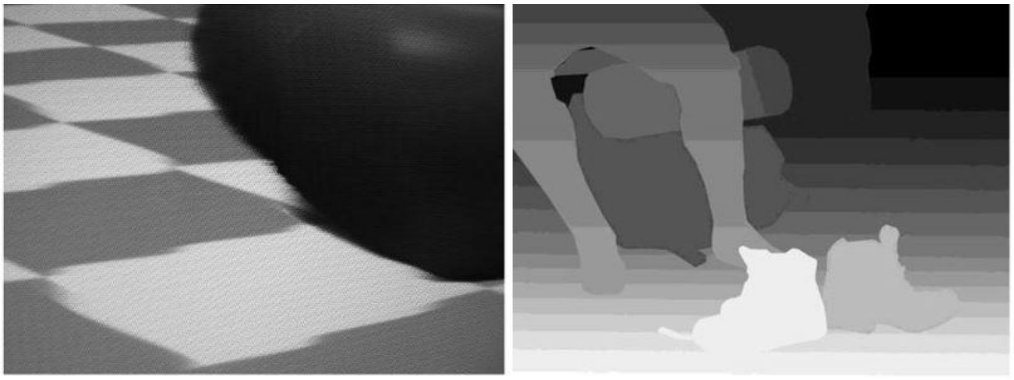


Figure 10: Depth map compression artifacts: a) depth ringing (left), card board effect (right) [7]

3.5.3 Artifacts in H.264 due to MVC: The artifacts can also happen in H.264 due to backwards compatible way of coding (third approach of MVC), i.e., interview prediction, where similarity of many views at any moment is used by MVC. This approach may favor in the increase of MC mismatch artifacts, which needs further investigation [7, 30].

3.6 Scalable video coding: Scalable video coding (SVC) is the encoding of high quality video having one or more sub streams, i.e., layered coding, etc., which is created by dropping packets in order to reduce the bandwidth which is required by the sub stream. Scalable video coding (SVC) refers to decoding required parts of a bit stream to produce smaller frame rate, spatial resolution, or quality [12, 32].

3.6.1 Artifacts due to temporal scalability: Temporal scalability is defined in terms of frame rate. No new artifacts are generated other than mosquito noise, MC mismatch, etc. The quantization parameter is increased in order to avoid pumping artifacts in higher temporal layers [23, 24].

3.6.2 Artifacts due to spatial scalability:

Spatial scalability is defined in terms of spatial resolution. The data which is decoded at lower resolutions can be used to reduce the bit rate and to predict the higher resolutions. Although the basic concept is same, but its implementation is different in MPEG-2 video, MPEG-4 Part 2 and H.264 SVC. In H.264, up sampling possibly increases blocking artifacts, as the block partitions of the lower layer are up sampled accordingly. Up sampling also enhances motion compensation error as same prediction area is used for MC due to the scaling of motion vectors and reusing the reference lists. Moreover, up sampling makes mosquito noise more visible as higher quantization parameters are used in the enhancement layer. Inter-layer intra prediction up sampling may also produce blurring due to bilinear filtering and favors mosquito noise due to the absence of high frequency coefficients [1, 7, 33].

3.6.3 Artifacts due to quality scalability: This scalability is a special case of spatial scalability as the video streams which are generated can be used to predict and decode the video with different qualities. Scalability relies on layers and quality scalability refers to the quality or SNR scalability in terms of reliability or combination of both. Coarse grain quality scalability uses same approach as used by inter layer inter prediction and produce similar artifacts. A drift is also introduced between encoder and decoder if enhancement layers are removed in quality scalability process, which is different from MPEG-2 codec drift, but needs investigation whether it introduces any new artifact or not [1,7, 33].

3.7 Artifacts due to new intra prediction modes in H.265: A new intra prediction mode, i.e., angular intra prediction, has been proposed in H.265 which may increase the probability of pumping artifacts further, which needs further research [2].

3.8 Artifacts due to interpolation filter in H.265: Similarly, in H.265/HEVC, the interpolation filter for subsamples is changed besides the change in transform size, i.e., a 6-tap directional or a 12-tap DCT based interpolation filter, as compared to Wiener and bilinear filter used in H.264. However, signal characteristics are changed in the interpolated subsamples correspondingly, and may expose new artifacts which need further investigation [2, 12].

The above section has comprehensively discussed the new artifacts produced due to the new coding tools and also their effect on classical artifacts. Table 3 shows a summary of these artifacts [1, 4, 7, 9].

Table 3: Summary of new artifacts (spatial & temporal) due to use of new video coding tools

Effect due to new coding tools/Artifact	Impact/Occurrence reason	Relation with other artifacts
Effect of new transforms and transform sizes	1.Integer transform used in H.264 & H.265 produce less transform coefficients as compared to DCT avoiding the loss of signal energy	1.Transform size of 4x4 in H.264/H.265 reduce ringing 3.Smaller transform sizes (4*4) in H.264 & H.265 may increase blocking distortion as the number of block borders increases
Effect of macro block partitioning	1.Enables to perform a separate search for each part of a macro block for fine matching	1.Without de-blocking, partitioning not only increases the probability of blocking but also enhances the appearance of MC (enhanced MC mismatch favors the appearance of mosquito noise)
Effect of larger size macro block in H.265	Coding efficiency increases as the macro block size increase	1. Macro block sizes of 32x32 and 64x64 in H.265 increase the probability of ringing artifacts

Effect of different coding modes	Due to the change of coding modes of one frame area in subsequent frames	Produce flickering or pumping artifact
Depth ringing or depth bleeding (MVC)	Quantization of depth maps yields depth ringing or depth bleeding; most prominent at steep edges	
Card board or puppet theater effect (MVC)	Both depth estimation and harsh quantization may also produce this effect	
Frame packing effect (MVC)	1.Frame packing is stereoscopic video coding (MVC), which can cause cross talk artifacts	1.Color bleeding propagates across views when using checkerboards arrangements, i.e., frame packing. 2.Mosquito noise, pumping and MC mismatch enhances due to interleaving of views. 3.Using frame alternation arrangements, MC mismatches increases
Artifacts in H.264 due to MVC	Artifacts can also happen due to backwards compatible way of coding	1. MC mismatch may increase due to the backwards compatible way of coding, which needs further investigation.
Artifacts due to temporal scalability	Temporal scalability is defined in terms of frame rate.	1. No new artifacts originate other than mosquito noise, MC mismatch, etc. 2. Quantization parameter is increased to avoid pumping artifacts in higher temporal layers. 3. Dropping of fine temporal resolution layers might create jerkiness.
Artifacts due to spatial scalability	Although the basic concept is same, but its implementation is different in MPEG-2, MPEG-4 Part 2 and H.264 SVC	1. Up sampling possibly increases blocking artifacts. 2. Up sampling favors MC mismatches 3. Up sampling makes mosquito noise more visible 4. Up sampling may also produce blurring due to bilinear filtering.
Artifacts due to quality scalability	Quality scalability relies on layers and is defined based on the quality or SNR scalability in terms of reliability or combination	1.Coarse grain quality scalability produce similar classical artifacts
Artifacts due to angular intra prediction mode in H.265	A new angular intra prediction coding mode is used in H.265	May increase pumping artifacts

In the previous sections, we have presented classical and new artifacts (spatial and temporal) comprehensively. Based on the above discussion, the next section proposes recommendations which can help in developing more efficient codec's and dealing with these artifacts more effectively.

4. Recommendations:

4.1 The study has shown that artifacts are generally compensated at the decoder side which may create other visual distortions. However, it would be the best option to develop such codec's which are themselves aware of different kinds of artifacts. In this way, if the sources of these new artifacts are known, then these artifacts can be avoided on encoder side rather than compensating at decoder side. For example, modifying the encoder is best option in order to reduce MC mismatch significantly rather than compensating [1, 7].

4.2 By the development of artifact aware encoder's and also taking into account HVS perception in encoder design, research focus will be on avoiding any artifact rather than compensation. Moreover, this awareness will reduce the post processing issues by the use of information provided by the encoder. In addition, new metrics can be developed which can make artifact detection easier and also apply compensation algorithms more selectively on the encoder and decoder side [1, 7].

4.3 Many codec's perform rate distortion optimization (RDO) to optimize the encoding. HVS effect should be taken into account while performing such calculations [1, 7].

4.4 New video quality algorithms should be developed, especially for most annoying artifacts and for new artifacts generated by new video coding tools.

4.5 New video coding tools (SVC, MVC, and NVC) should be analyzed in detail to find out that how they affect the existing artifacts and can produce new artifacts [2, 12].

4.6 Existing video quality algorithms should be improved in efficiency so that each one of them are quite efficient and reliable to detect classical spatial/temporal and new artifacts

4.7 Pre- and post-processing techniques should also be enhanced and developed to eliminate/minimize the effects of artifacs and enhance the efficiency of image/video quality evaluation.

4.8 Effect of one artifact on other artifacts, i.e., masking/superposition, creation, decrease, etc., needs more in depth analysis, especially the effect of new artifacts on classical artifacts.

4.9 There are many other existing video coding tools whose effects on classical and new artifacts must be analyzed in depth.

4.10 Different artifacts and their effect on HVS need further research as human is final observer of video quality.

4.11 Existing literature generally focus on spatial artifacts. In the same way, temporal artifacts should be analyzed in more detail, especially in new coding tools.

4.12 Video coding standards generally define video decoders only, but encoder configurations are generally different. However, more uniform kind of encoders should be designed to reduce or eliminate the artifacts.

As the blocking distortion is most annoying and it is also present in all classical and new coding tools, therefore, the next section proposes a new novel NR blockiness distortion algorithm for an accurate measurement of blockiness.

6. Conclusion

With the advent of new video coding tools which provide a decreased bit rates compared to previous standards, new spatial and temporal artifacts have emerged along with the classical artifacts. There is still a lack of awareness about classical and new artifacts, especially due to the use of new coding tools. Therefore, this research provides a comprehensive overview of the spatial and temporal artifacts produced by current and new video coding tools, i.e., MPEG-2 Video, MPEG-4 Part 2, H.264, VC-1, SVC (scalable video coding) and MVC (multi view coding) by H.264, and H.265/HEVC. Many existing papers in the literature discuss these artifacts, but none of the paper comprehensively describes all these artifacts as presented in this study. The paper also discusses artifacts produced due to the new coding tools and also their effect on classical artifacts due to these new coding tools.

The first contribution of the paper is a comprehensive survey and analysis of the classical spatial and temporal artifacts as compared to the existing literature. The summary has also been provided for these spatial and temporal artifacts which does not exist in the current literature. The summary further shows that how these artifacts are produced and related to each other, etc. Another contribution of the paper is a comprehensive analysis/survey of the new artifacts (both spatial and temporal) produced due to the use of SVC (scalable video coding) and MVC (multi view coding) by H.264 and H.265/HEVC, respectively. A summary of these artifacts has also been provided which highlights the relation of these artifacts, i.e., how these artifacts are produced and related to other artifacts, etc. Another contribution of the paper is the discussion of the effect of new coding tools on classical artifacts which is not discussed in most of the papers.

Several recommendations have also been proposed as a contribution towards the paper based on the detailed analysis of the artifacts. These recommendations highlights the future work with respect to the knowledge and understanding of the new artifacts produced due the new coding tools, i.e., SVC, MVC, H.265/HEVC, etc., and their effects on the classical artifacts. The artifacts produced due to the new coding tools, especially needs more in depth research.

The research has also several key understandings/observations, i.e., sources of the existing and new artifacts should be known, etc. The artifacts aware encoders can be developed with this knowledge which is more efficient rather than compensating the artifacts at the decoder side. This research can be used in any image processing applications where received quality of the image/video is not up to the standard and needs further investigation, especially to see the effect of one artifact to the others.

7. References

[1] Richardson, I.E.G. (2003) H.264 and MPEG-4 Video Compression. Chichester, England: Wiley & Sons Ltd.

[2] Sullivan, G., Ohm, J., Woo-Jin, H., and Wiegand, T., "Overview of the High Efficiency Video Coding (HEVC) standard," IEEE Trans. Circuits and Systems for Video Technology 22(12), 1649–1668 (2012).

[3] B. M. Zavodsky, "Time-variant video quality evaluation for mobile networks," Phd Thesis, Fakulta elektrotechniky a informatiky Katedra rdioelektroniky, Slovensk technick univerzita v Bratislave, 2006.

[4] Yuen, M. and Wu, H., "A survey of hybrid MC/DPCM/DCT video coding distortions," Signal Processing 70 (3), 247–278 (1998).

[5] WeisiLin, C. Jay Kuo, "Perceptual visual quality metrics: A survey", Journal of Visual Communication and Image Representation, volume 22, Issue 4, May 2011, pages 297-312.

[6] Muhammad Shahid, Andreas Rossholm, Benny Lövström, and Hans Zepernick, "No-reference image and video quality assessment : a classification and review of recent approaches", EURASIP Journal on Image and Video Processing, 2014, pages 397-437.

[7] Andreas Unterweger, "Compression artifacts in modern video coding and state of the art means of compensation", University of Salzburg, Austria, 2012.

[8] Chodisetti Rakesh Anil, "H.264 video coding artifacts; Measurement and reduction of flickering ," Master thesis, Department of Electrical Engineering, Blekinge Institute of Technology, 2013.

[9] K. Zeng, T. Zhao, A. Rehman, and Z. Wang, "Characterizing perceptual artifacts in compressed video streams," IS&T/SPIE Human Vision and Electronic Imaging, vol. 9014, no. 10, pp. 2458–2462, Feb. 2014.

[10] Liqun Lin, Shiqi Yu, Tiesong Zhao, Member," PEA265: Perceptual Assessment of Video Compression Artifacts", Computer Vision and Pattern Recognition; Image and Video Processing, 1601–1612 (2019).

[11] Zeng, K., Rehman, A., Wang, J., and Wang, Z., "From H.264 to HEVC: Coding gain predicted by objective video quality assessment models," in International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Feb 2013.

[12] Hanwha Techwin, "White Paper: H.265 High Efficiency Coding Video Compression for Security Applications", Ridgefield Park, NJ, USA, 2016

[13] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 4, pp. 533–546, April, 2015.

[14] S. Yoo, K. Choi, and J. Ra, "Blind post-processing for ringing and mosquito artifact reduction in coded videos," IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 5, pp. 721–732, May 2014.

[15] H Liu, N Klomp, I Heynderickx, "A no-reference metric for perceived ringing artifacts in images", IEEE Trans. Circuits Syst. video technol. volume 4, issue 20, pages 529–539, October, 2010.

[16] N. Suresh, "Mean time between visible artifacts in visual communications," Phd Thesis, Department of Electrical and Computer Engineering, Georgia tech. University, 2007.

[17] C. Dong, Y. Deng, C. Chang, and X. Tang, "Compression artifacts reduction by a deep convolution network," in Proc. IEEE International Conference on Computer Vision (ICCV), Feb. 2016, pp. 576–584.

[18] J. Xia, Y. Shi, K. Teunissen, and I. Heynderickx, "Perceivable artifacts in compressed video and their relation to video quality," Signal Process Image., vol. 24, no. 7, pp. 548–556, Aug. 2009.

[19] M. Rohil, N. Gupta, P. Yadav, "An improved model for no-reference image quality assessment and a no-reference video quality assessment model based on frame analysis", Signal, Image and Video Processing, vol 6, no.7, pp. 610-619, July 2019.

[20] J. Liu , D. Liu, W. Yang , S. Xia , X. Zhang , and Y. Dai," A Comprehensive Benchmark for Single Image Compression Artifact Reduction", IEEE Transactions on image processing, vol. 29, pp. 225-241, 2020

[21] K. Yang ; S. Wan ; Y. Gong ; Y. Yang ; Y. Feng, "Temporal Distortions in H.265/HEVC", Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, South Korea, 13-16 Dec. 2016.

[22] Y. C. Gong, S. Wan, K. Yang, F. Z. Yang, and L. Cui, "An efficient algorithm to eliminate temporal pumping artifact in video coding with hierarchical prediction structure," J. visual communication, vol. 25, no. 7, pp. 1528–1542, Oct. 2014.

[23] Y. Yang, J. Xiang. and H. Wau., "Robust filtering technique for reduction of temporal fluctuation in H.264 video sequences," IEEE Trans. Circuits and Systems for Video Technology 20(3), 458–462, Nov, 2010.

[24] J. Lee, J. Park, H. Wang, "A New Distortion Measure for Motion Estimation in Motion-Compensated Hybrid Video Coding. Signal Processing", Image Communication, vol. 2, no. 6, pp. 75–84, 2011.

[25] P. V. Pahalawatta and A. M. Tourapis, "Motion estimated temporal consistency metrics for objective video quality assessment," in Quality of Multimedia Experience, International Workshop on, pp. 174–179, 2009.

[26] S. Wan, Y. Gong, and F. Yang, "Perception of Temporal Pumping Artifact in Video Coding with the Hierarchical Prediction Structure," in Proc. IEEE Int. Conf. Multimedia and Expo, pp. 503-508, Jul. 2012.

[27] Y. Gong, S. Wan, K. Yang, B. Li, and H. Wu, "Perception based Quantitative Definition of Temporal Pumping Artifact", in Proc. of the 19th Int. Conf. on Digital Signal Processing, pp.870-875, August 2014.

[28] Y. Qi and M. Dai , "The Effect of Frame freezing and Frame skipping on video quality", Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Jakarta, Indonesia, 13-16 April, 2006.

[29] Z. Akhtar and T. Falk, "Audio-Visual Multimedia Quality Assessment: A Comprehensive Survey", IEEE access, vol. 11, no. 7, pp. 412-440, July, 2017

[30] G. Vetro, A. Wiegand, T. Sullivan, "Overview of the Stereo and Multiview Video Coding Extensions of the H.264/AVC Standard". Proceedings of IEEE, Special Issue on 3D Media and Displays, vol. 99, no. 4, pp. 626–642, 2011.

[31] P. Merkle, Y. Smolic, P. Wiegand, "The effects of multiview depth video compression on multiview rendering", Signal Processing: Image Communication, vol. 24, no. 1, pp. 73–88, 2008.

[32] H. Boujut, B. Pineau, J. Hadar, O. Ahmed, "Weighted-MSE based on saliency map for assessing video quality of H.264 video streams", Proceedings of the SPIE "Image Quality and System Performance VIII", vol. 7, no. 6, pp. 686–696. 2011.

[33] R. Farrugia, A. Debono,"A Hybrid Error Control and Artifact Detection Mechanism for Robust Decoding of H.264/AVC Video Sequences", IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 5, pp. 756–762, 2010.

An Overview of Social, Economic, Environmental, and Safety Impacts of Intelligent Electric Vehicles

Dr. Muhammad Uzair,
Faculty of engineering, Islamic University of Madinah, KSA
uzair91@hotmail.com, muzair@iu.edu.sa
and
Saleh Hosain,
Faculty of engineering, Islamic University of Madinah, KSA
salehalhafiz@hotmail.com

Abstract: Internet of Things (IoT) can connect millions of devices, objects, appliances, sensors, and applications together to collect huge amount of data for multipurpose applications. It requires an efficient network to carry out these processes in a faster way which is accomplished with the help of 5G networking, i.e., ability to provide a very high data rate, etc. Autonomous vehicles (smart cars) are one of the applications of the IoT which integrates its components to make smart cars working as robots. Smart cars will have great impact on our society (like different driving experience, improving accessibility to urban opportunities to different groups of people, providing a convenient, comfortable, and flexible transportation, etc.), environment (like fuel consumption, gas emissions, land use, and urban spatial structure, etc.), economy (many current practices will close with new opportunities), and safety/security (software failures, denial of service, eavesdropping, hijacking, etc.). These impacts will be profound, particularly for level 5 automation smart cars.

This research analyzes key aspects of the autonomous vehicles with respect to these issues along with ethical dilemmas. This study clearly shows that lots of work still needs to be done by the manufacturers, the vendors, the industry, and the governments in order to make this technology adaptable. The study also describes that until now many aspects of this technology are also ambiguous, i.e., technology will have positive or a negative impacts, which group of the society will get more benefits, and which group might be affected negatively, etc.

Therefore, the study highlights the need of more empirical results from industry, and research community in order to make a conclusion of the impacts of the technology. Similarly, greater emphasis should be given to the safety and security because if this technology is not a hundred percent safe, it will not be widely accepted. An attack on a smart car does not only threaten the safety and privacy of drivers, passengers and everyone traveling on the road, but also has a big impact on the auto industry, i.e., millions of cars have already been recalled in many cases. Finally, the research also highlights different challenges/issues faced by this technology. The research also proposes recommendations and solutions in order to handle these current and future challenges keeping in view that the security/safety is one of the biggest challenges for the autonomous vehicles (AVS). At the end, the research also proposes a new comprehensive layered architecture for the AVS with particular focus towards security and safety of the AVs and making it safer and reliable as compared to the existing architecture.

Keywords: Autonomous vehicles, safety/security, social, economical, environment

نظرة عامة على التأثيرات الاجتماعية والاقتصادية والبيئية المتعلقة بسلامة السيارات الكهربائية الذكية

الملخص: يمكن لإنترنت الأشياء (IoT) توصيل ملايين الأجهزة والأشياء وأجهزة الاستشعار والتطبيقات معًا لجمع كمية هائلة من البيانات للتطبيقات متعددة الأغراض. يتطلب هذا شبكة فعالة لتنفيذ هذه العمليات بطريقة أسرع والتي يتم إنجازها بمساعدة شبكات الجيل الخامس (5G)، أي القدرة على توفير معدل بيانات مرتفع للغاية، وما إلى ذلك. المركبات المستقلة (السيارات الذكية) هي أحد تطبيقات إنترنت الأشياء الذي يدمج مكوناته لجعل السيارات الذكية تعمل كروبوتات. سيكون للسيارات الذكية تأثير كبير على مجتمعنا (مثل تجربة القيادة المختلفة، وتحسين إمكانية الوصول إلى الفرص الحضرية لمجموعات مختلفة من الناس، وتوفير وسيلة نقل عملية ومريحة ومرنة، وما إلى ذلك)، والبيئة (مثل استهلاك الوقود، وانبعاثات الغاز، واستخدام الأرض، والهيكل المكاني الحضري، وما إلى ذلك)، والاقتصاد (سيتم إغلاق العديد من الممارسات الحالية بفرص جديدة)، والسلامة / الأمن (فشل البرامج، ورفض الخدمة، والتنصت، والاختطاف، وما إلى ذلك). ستكون هذه التأثيرات عميقة، خاصة بالنسبة للسيارات الذكية ذات التشغيل الآلي من المستوى الخامس.

يحلل هذا البحث الجوانب الرئيسية للمركبات ذاتية القيادة فيما يتعلق بهذه القضايا إلى جانب المعضلات الأخلاقية. تظهر هذه الدراسة بوضوح أنه لا يزال يتعين القيام بالكثير من العمل من قبل المصنعين والبائعين والصناعة والحكومات من أجل جعل هذه التكنولوجيا قابلة للتكيف. توضح الدراسة أيضًا أنه حتى الآن هناك العديد من جوانب هذه التقنية غامضة أيضًا، أي أن التكنولوجيا سيكون لها تأثيرات إيجابية أو سلبية، وأي مجموعة من المجتمع ستحصل على المزيد من الفوائد، وأي مجموعة يمكن أن تتأثر سلبًا، وما إلى ذلك. تسلط الدراسة الضوء على الحاجة إلى المزيد من النتائج التجريبية من الصناعة ومجتمع البحث من أجل التوصل إلى استنتاج بشأن تأثيرات التكنولوجيا. وبالمثل، يجب التركيز بشكل أكبر على السلامة والأمن لأنه إذا لم تكن هذه التكنولوجيا آمنة بنسبة مئة بالمائة، فلن يتم قبولها على نطاق واسع. إن الهجوم على سيارة ذكية لا يهدد فقط سلامة وخصوصية السائقين والركاب وكل من يسافر على الطريق، بل له تأثير كبير على صناعة السيارات، أي أنه تم بالفعل استدعاء ملايين السيارات في كثير من الحالات. أخيرًا، يسلط البحث الضوء أيضًا على التحديات / المشكلات المختلفة التي تواجهها هذه التكنولوجيا. يقترح البحث أيضًا توصيات وحلولًا للتعامل مع هذه التحديات الحالية والمستقبلية مع الأخذ بعين الاعتبار أن الأمن / السلامة هو أحد أكبر التحديات التي تواجه المركبات ذاتية القيادة (AVS). في النهاية، يقترح البحث أيضًا بنية جديدة شاملة متعددة الطبقات لنظام AVS مع التركيز بشكل خاص على أمن وسلامة المركبات وجعلها أكثر أمانًا وموثوقية مقارنة بالبنية الحالية.

Introduction: Internet of things (IoTs) is a communication technology, which makes everything capable of communicating with other things existing in this world such as homes, cars, animals, humans, farms, and industry, etc. It uses low cost processors, i.e., Raspberry Pi and Arduino microcontroller, different protocols, and wireless networks to add digital intelligence in these dumb devices so that these devices can communicate with each other. In this way, it basically merges the digital and physical world [1]. In order to enable this merging and optimize the energy consumption, coverage, data rates, and security it requires next generation (5G) of the mobile networking system. Therefore, 5G is considered a driver for the Internet of Things (IoT) to connect different devices, sensors, and things together for data collection, and further processing [2].

Autonomous vehicles (AVs), or smart cars are one of the most trending researches in today's automotive industry. Smart cars are one of the great applications of the IoT which has the ability to connect the internet. Smart cars have the ability to send and receive data onto internet with other devices, both inside and outside of the car. The smart car is like a robot that uses its own intelligence to perform different functions. It encompasses cases such as telematics (i.e., fleet management or Geo-fencing), connected infotainment such as access to drive information (i.e., speed), and control functions (i.e., air conditioning), etc. Vehicle-to-Infrastructure (V2I) communication technologies will enable the drivers/cars to get information from the control station regarding parking, speed cameras, and real time information about traffic, etc. Vehicle-to-Vehicle (V2V) communication technologies will enable the drivers/cars to share information about each other regarding location, collision avoidance, traffic streamlining, etc. [3]. Figure 1 shows that how smart cars work [4].

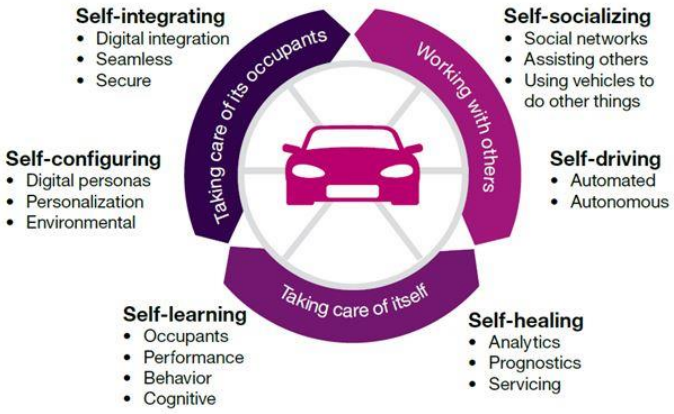


Figure 1. Smart Car [4]

The overall structure of the research is shown in Figure 2.

Levels of automation: Smart cars have two levels of automation, i.e., level 0 to 3 (low level automation), and level 3 to 5 (high level automation). Each of these automations provides an entirely different driving experience (especially level 5) by providing a completely different architecture and interfaces as compared to the today's cars. Table 1 shows the standard levels of automation [4].

Table 1: Standard levels of automation in AVs

SAE Level	Definition	System capability
0	No Automation. Drivers perform all tasks	N/A
1	Driver Assistance. Vehicles assist some functions, but drivers handle accelerating, braking, etc	Some driving modes
2	Partial Automation. Vehicles can assist with steering or acceleration/deceleration functions, etc.	Some driving modes
3	Conditional Automation. Vehicles control all monitoring/functions, but drivers respond to request	Some driving modes
4	High Automation. Vehicles control all monitoring/functions even if drivers do not respond appropriately	Many driving modes
5	Full Automation. No human attention needed for any condition that can be managed by a human driver	All driving modes

Expectations from smart cars: In today's digital life, people have experiences of digital interfaces in every part of life, i.e., customer services, retail, etc. People are always interested in how to feel the use of interact with a product and about the company which provides it. Similarly, people who will use the smart cars in the near future, i.e., drivers, or passengers will have their own opinions towards the making, model, price, or the appearance of a vehicle. Both physical and digital developments will ultimately make the opinion of the driving experience. Therefore, the car-makers will need to develop such services and applications which offer the best personalized driving features. Software advancements will be the main factor in determining the popularity of certain vehicles over the others. With increasing understanding among the car-makers, developing reliable and user specific software in the smart cars will provide an opportunity to get more shares in the auto market [5]. Similarly, humans are giving the control of their lives to a robot, i.e., smart cars. This phenomenon is only possible when there is a good deal of trust which can only be achieved through reliability, security, and safety. Some recent accidents show that there are many severe shortcomings in the smart cars, and

these smart cars are not trustworthy for a safer drive under all weather conditions. The safety and security solutions to level 5 automation are particularly important as it will have cutting edge effects on the society in every aspect. The concept of smart cities will also not be completed without smart cars. For a smart city, smart cars should have the ability to move through intersections without any delay, immediately self report of any accident, make ways for the emergency vehicles, and should know where to park safely, etc. [6].

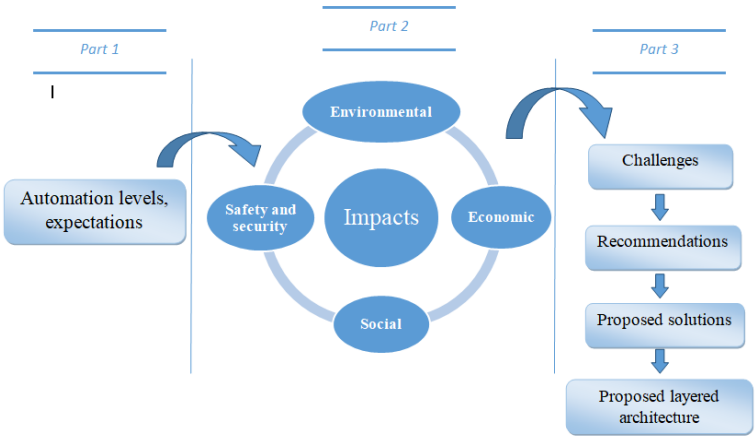


Figure 2: Overall structure of the research

The remainder of the paper is organized as follows: Section I, and section II describe the environmental and economic impacts of the autonomous vehicles. Section III, and section IV present the safety and social impacts. Section V presents the challenges to the autonomous vehicles. Section VI presents the recommendations and section VII presents solutionsto handle the existing and future challenges. Section VIII presents a new comprehensive layered architecture for the AVs. Section IX presents the conclusion.

In the following section, the environmental impacts of the AVs are being presented.

1. Environmental impacts

1.1. Radiation effects on humans, animals, and plants: With the rapid increase in the wireless technology and IoT, there will be an increasing number of devices around us, which will radiate high amount of radio frequency (RF). The smart cars will also be responsible for wireless radiations as they comprise of lots of numbers of IoT devices. In a recent study, the World Health Organization has classified wireless radiation as a Group 2B, i.e., possible human carcinogen [7]. The studies have also found that exposures to radio frequency radiation can also cause disruption of normal brain development in fetuses, learning disabilities, heart

abnormalities, and electro hypersensitivities. The pregnant women, children, people with implanted medical devices, electromagnetically sensitive individuals, and the elderly people will particularly be affected by these radiations. Similarly, the sweat ducts within our skin act as an antennae. Due to this, the millimeter waves can penetrate up to 1 to 2 millimeters into skin tissue, and are also absorbed by the surface layers of the eye's cornea. The electromagnetic radiations also damage trees, and change the structures of the plants. The research studies also indicate that the wireless technology is another contributory factor to the decline birds, frogs, bats, and honeybee populations [8].

1.2. Gases emissions: Smart cars will have both positive and negative effects on the environment, i.e., motor vehicles can reduce transport energy consumption up to 90 percent, or increase it by more than 200 percent. According to the Environmental Protection Agency (EPA), the transport sector is at least responsible for a quarter of greenhouse gas emissions. Drivers will receive the advance information about traffic congestion by using smart cars, which will enable people to reschedule, postpone, cancel, or make their journey creating a fuel efficient and environmentally friendly environment. AVs which are powered electrically will reduce the carbon dioxide emissions. This would be a major positive effect on the environment, but it will only be true when we use the green energy sources of electricity production [9]. The net emissions are likely to reduce, but it is not clear that how much it will reduce. On the other hand, the total number of miles will be increased due to easy travel, and will encourage people to make an extra trip causing negative effects on the environment. According to the Eno-Center study, the emissions per mile could decline (if AV communicates well with the infrastructure) even if the vehicle mileage travel (VMT) increases. Therefore, AVs will have a significant impact on the environment, but good or bad will depend on the technological and policy options that are still evolving and are not certain [10-12].

1.3. Infrastructure: One of the most important limiting factors in the adoption of autonomous vehicles is the infrastructure. It will be easier to run the AV in any area if it is more understandable and predictable, i.e., traffic environment. This thing will also make AVs quickly adapted by the society. Therefore, the investment in the infrastructure will be a significant factor that will determine the success of AVs in the future. Moreover, when more numbers of AVs will be on the road; this technology will improve the infrastructure which will turn into a package of great societal benefits. This is particularly possible in highly congested areas where parking spaces are expected to disappear as AVs do not need to stay at the customer destination, i.e., either AV can return to the specified places or roam through the city. An

average vehicle is parked 95 percent of the time and the lot of expensive urban space is dedicated to parking, especially in the busy areas. The study also shows that a reduction of 90 percent in parking areas is possible due to this technology [10].

1.4. Noise and light reduction: Electric engines can reduce the noise emissions at lower speeds. This noise reduction may increase the building of roads near the residential areas thus reducing the travel distance. But on the other hand, the cost of the land close to the highways may also be increased. Similarly, AVs depend on sensors which may not need light to see the surroundings as human drivers require. Therefore, it will be possible to build tunnels and similar structures without lights to reduce the light emissions and save electrical energy [12-13].

1.5. Fuel efficiency: The studies by the Eno Center for transportation have derived that self driving car technology could reduce fuel consumption of the cars by regulating the driving manner since everything is automated. The study finds that fuel economy from 23 to 39 percent is possible with smooth traffic, i.e., by minimizing braking on the highways, which is possible by integrating AV technology with V2V and V2I communications. This will also depend on the efficiency of the traffic smoothing algorithms. However, these and other analysis still do not have an exact measurement that how much fuel consumption will be achieved as it will really depend on the market share of this technology and number of other factors [12-14].

1.6. Congestion mitigation: AVs may reduce the energy wastage of the traffic by improving the traffic flow, and consequently also reduce accidents frequency. Schrank estimated the fuel wasted in the U.S every year due to traffic since 1982, and found that the fuel wasted due to traffic congestion rose steadily from 0.5 percent to 1.8 percent from 1984 to 2005, and it is expected to reach 2.6 percent by the end of the next year, i.e., 2020. Based on this study, if other factors remain the same, then approximately 4.2 percent of the fuel would be wasted due to congestion till 2050. Overall, a complete elimination of the congestion by using smart cars might decrease the energy intensity by about 2 percent today, and a little over 4 percent till 2050 [12].

1.7. Automated Eco driving: Eco driving teaches us how to reduce the fuel consumption without changing the vehicle design. The energy consumption can be reduced by running the engine at its most efficient operating points, i.e., high load and moderate speed. Another practice is to minimize braking as braking increases the energy wastage. Similarly, training the drivers is also the part of Eco driving. The study showed that driver's training, i.e., teaching optimal acceleration and deceleration brought a reduction in the energy consumption from 0 to

26 percent. However, without regular reminders, the drivers who took an Eco driving course stopped following efficient ways with the passage of time. Another study found that average savings are about 20 percent in the short run and about 10 percent in the long run. Similarly, another study found that in heavily congested conditions energy consumption can be reduced about 35 to 50 percent by using optimal driving. However, a driving cycle has to optimize keeping in view the legal constraints (e.g., speed limit) and maintaining the required travel time. More energy consumption can be seen especially in urban areas which have more frequent stops. AVs have the ability of Eco driving due to their ability to acquire current traffic and road conditions [12].

1.8. Impact on city planning: The increased automation will change the phenomena of planning and building the cities in the future. There will be removal of most of the traffic signs, lights, and parking spaces, etc. But on the other side, more parking spaces are needed for picking up, and dropping off customers especially during any public event. Therefore, connected cities will emerge by managing traffic data for the efficient flow of the traffic, i.e., intelligent traffic lights working with the current demand on the roads, automated crossings with no traffic lights, etc. There will also be mobile car charging facilities distributed all around the city, and/or charging the AVs by using electric field Street while driving [13].

1.9. Clean vehicle technologies: The reduction in the accidents will provide a cushion to the manufacturers to build lighter vehicles. With light weight vehicles, there is a possibility of weight reduction up to 20 percent in the vehicles, and every 10 percent reduction corresponds to a six to seven percent reduction in the fuel consumption. This reduction in weight can be further used to improve batteries or hydrogen fuel cells for clean vehicle technology, i.e., Nissan Leaf battery weighs 600 pounds. Therefore, with alternate fuel vehicles, i.e., able to self-charge, large climate benefits are possible [14].

1.10. Platooning: Running the vehicles together closely is called platooning, which also helps to achieve fuel efficiency due the cut down on air drag resistance. This is especially attractive for freight trucking on highways for long distances, which may be called as *road trains*. The total fuel efficiency can be from 10 to 20 percent depending on the type of vehicle involved, trip distances [14].However, tight vehicle spacing on roads could cause problems for other motorists.

1.11. Disposal of batteries: Lithium-ion battery waste management would be a great task for the manufacturers also. There should be possibility of re-using or recycling of components of batteries and other manufactured components for electric battery vehicle technologies to

reduce the environmental impacts. It would depend on the technology advancement and advancements in membrane and fuel storage technologies [13].

1.12. Travel behavior patterns: The change in the travel behavior due to the AVs might have more effect on environment as compared to social or economical. The greenhouse gas emission might decrease due to the AVs on a functional unit bases, but, it may increase due to the more vehicle miles travelled [13].

Table 2 shows a high-level summary of the environmental impacts (qualitative) due to the AVs.

Table 2: High level summary of environmental (qualitative) impacts

Impact	Scale	Importance	Outcome	Advantages/Disadvantages
Gases emissions	Big	High	An overall decrease of the gases emissions	Reduction of the gases emissions, especially if AVs use other renewable energy sources
Congestion mitigation	Small	Medium	Possible decrease in traffic in big cities	Reduction in energy wastage by improving the traffic flow
Infrastructure	Medium	High	Increase in the overall required infrastructure	Extra money requiredto improve the current infrastructure or build new infrastructure
Fuel efficiency	Small to medium	High	Reduction of the use of fossil fuels	Advanced technology of electric machine can have more efficiency
Clean vehicle technologies	Medium	Medium	Ambiguous	Slow charging of cars by using clean energy sources, i.e., solar energy & wind, etc.

Tables 3 and 4 show a high-level summary of the environmental impacts (quantitative) due to the AVs [15].

Table 3: High level summary of environmental (Quantitative) impacts [15]

Aspect	Effect magnitude	Time	Notes
Fuel economy	+100% - +1000%	By 2050	Can be achieved by vehicle weight reduction
Fuel demand	-91% - +173%	90% AVs penetration	Based on different cases taking to account all possibilities
VMT	+9%	90% AVs penetration	That is also given for low market share
Fleet size	-42.6%		
Energy use	-12%	Fleet is all shared AVs	AV compared to the average light duty vehicle
GHG	-5.1%		
Fleet size	-66%	Fleet is all shared AVs	Does not depend on Energy only output

Table 4: High level summary of environmental (Quantitative) impacts [16]

Aspect	Decrease	Little decrease	No change	Little increase	Increase	Uncertain
VMT	1.3%	12.1%	17.8%	28.7%	29.3%	8.9%
Energy use	3.2%	41.0%	16.7%	28.2%	3.2%	7.7%
Walking and biking	0.0%	14.6%	53.5%	26.8%	0.6%	4.5%
Congestion	4.4%	34.2%	27.2%	19.0%	7.6%	7.6%
Pollution	5.8%	45.5%	26.9%	14.7%	1.9%	5.1%

Figure 3 shows a percentage of energy consumption changes due to the vehicle automation [12].

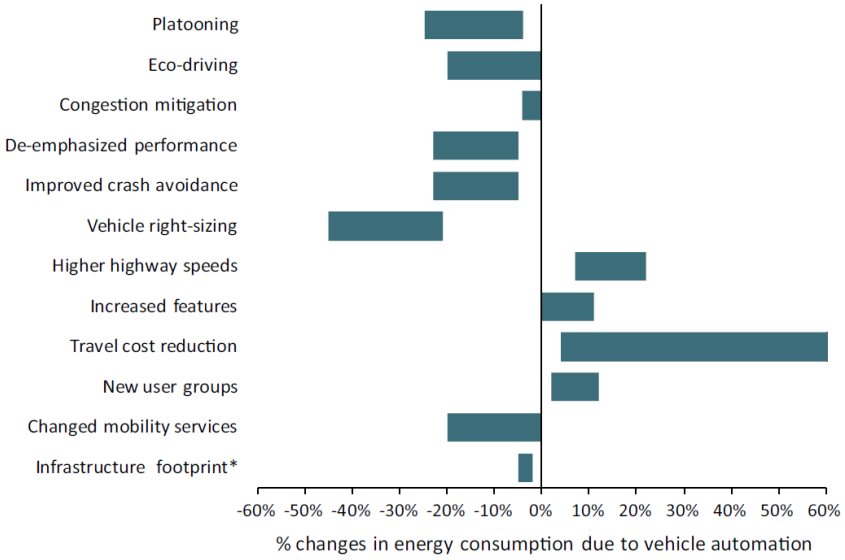


Figure 3: Percentage of energy consumption due to vehicle automation [12]

The economic aspects of the smart cars are described in the next section.

2. Economic impacts

2.1. Impact on car industry: Autonomous vehicles will have significant effects on auto industry due to the current developments, i.e., electric vehicles, increasing automation, artificial intelligence, car sharing, and on demand culture, etc. This technology will change the whole scenario that how cars will be manufactured, used, maintained, and how the legal issues (liability and insurance) along with the ethics will be answered. With this technology, the maintenance of the car would be done in traditional car workshops, or with a large scale service operator, or the car manufacturer has to provide the maintenance due to the complicated

technology. As AVs will be able to discover the failures by themselves and drive the cars to a workshop, it will also cause the closure of the small garages. Therefore, the auto industry needs to do more accurate work in order to maintain a stable economic situation and keep its lead in the industry; otherwise any new company will take over the market [17].

2.2. Impacts on logistics: The autonomous vehicles can have a big impact on the cost delivery of goods. It is estimated that the current supply of goods will increase at least three times till 2050. The wages of the drivers generally contribute up to 30 percent of the total cost of the supply. By using AVs, this amount will be saved with an efficient and on time delivery of goods as AVs do not need rest like drivers. However, the number of jobs will be reduced in logistics by using this technology [10].

2.3. Changes in vehicle miles traveled (VMT): The study shows that VMT per AV will be 20 percent higher than regular vehicles when there will be 10 percent market penetration and 10 percent higher when there will be 90 percent market penetration. The additional VMT increase may be realized if travel costs and congestion is reduced. However, Fagnant and Kockelman undermine this analysis [18]. Another study finds that a fleet of shared AVs in Texas, serving more than 56,000 trips a day, was found to travel 8.7 percent more in the empty area. This number reduced to 4.5 percent and less than 1 percent when the demand rose by a factor of 5 and ride-sharing was permitted. Similarly, various studies show that each AV could serve the same number of trips as 10 privately owned vehicles if all traveling lies within the 12 to 24 miles Geofence. Although, the ratio from 10 to 1, may be too high, especially in rural areas and areas with longer distances [18-20].

2.4. Revenue collection and vehicle ownership: Smart cars have brought a revolution in the auto industry. Business Insider expects that approximately 380 million smart cars will be on the road till 2020 bringing total revenue of around 8 trillion dollars to the governments. This revolution has also brought different changes in the business market finishing the traditional concepts such as taxi business, i.e., even a blind person will be able to communicate by using hundred percent drivers less cars. Similarly, the study shows that till 2030 car ownership will fall by 80 percent in the United States, as AVs can be owned and used by the individuals, the families, the investors, or the public collectively [20].

2.5. Discount rate and technology costs: The technology is evolving rapidly, but many things are still not certain even the cost. A study found ten percent discount rate with respect to the net present value calculation, which is even higher than the 7 percent rate required by the federal Office of Management and Budget (OMB) for federal projects and TIGER grant

applications in USA [19]. The research finds that an additional 10,000 dollars would be required to purchase AVs (around seven years) when there will be 10 percent market share. However, the research assesses that the price would be reduced to 3000 dollars when market penetration will reach up to 90 percent. The researcher also assumes that the return rate would be equal to 37, 500 dollars, which is closer to the added AV technology. Similarly, Shchetko (2014) finds that LIDAR systems are the most expensive parts of AVs. In order to make affordable, future AVs should use non-LIDAR sensors, or LIDAR prices must fall [18,20].

2.6. Impact on travel cost: The study finds that shared AVs can be operated at a price of 0.42 to 0.49 dollars per mile which is almost the same as the today's car sharing services. Similarly, another study finds that shared AVs with at least 2 passengers will be competitive with mass transit till 2035. The same study shows that AVs are expected to become cheaper than other means of transportation as technology will grow. This will change the customer approach towards traveling mode, i.e., cost awareness will increase as per distance or per duration approach will evolve. This will also increase the understanding of the costs of travel as compared to today [10, 19].

2.7. Vehicle cost: Cost is always an important factor and it depends heavily that how much will be the cost of using an AV. The study shows that the average willingness to pay (WTP) for AV technology is 7253 dollars from 0 to level 3 automation. People are ready to pay another 5551 and 4589 dollars for additional level 4 and 5 automation. Mosquet (2015) finds that people are also ready to spend another 5000 dollars to add AVs features in their regular cars. With advancement in the AV technology and large scale productions promises will make AVs more affordable over the period of time [18, 20].

2.8. New business opportunists: It is worth to talk about the idea of not buying a car. With smart cars, new concepts of businesses are penetrating in the market and they might destroy traditional concepts. One of these ideas is to rent a smart car for some time using an application. The user can select a car with its specifications, and the car will come to the house door. Then the user can simply leave the car anywhere and it will go back automatically. This will create a new business concept and finish the others, such as taxi or Uber driver [10].

2.9. Employment issues: Highly skilled labor will be required in science and technology to design, build, and continuously improve the technology. On the other hand, low skilled workers and routine labor in the private sector will be highly affected, i.e., drivers and technicians, etc.

Table 5 shows a high level summary of the economical impacts (qualitative) due to the AVs.

Table 5: High level summary (Qualitative) of economical impacts

Impact	Scale	Importance	Outcome	Advantage/Disadvantages
Auto industry	Big	High	Ambiguous	Ambiguous
Other businesses	Medium	High	Ambiguous	Possible opportunities of new business
Employment	Medium	High	Possible reductions of employments in different sectors	High technology jobs will increase
Vehicles miles traveled	Medium	Medium	Ambiguous	Ambiguous

Table 6 shows a high-level summary of the economical impacts (quantitative) due to the AVs [21].

Table 6: High level summary (Quantitative) of economical impacts [21]

Aspect	Industry revenue ($B/yr)	Industry Impact ($B/yr)	% change in industry	$ per capita per year
Automotive Industry	$570	$42	7%	$132
Electronics and software	$203	$26	13%	$83
Auto repair	$58	$21	36%	$66
Traffic police	$10	$5	50%	$16
Infrastructure	$169	$8	4.4%	$24
Freight transportation	$604	$100	17%	$313
Personal transportation	$86	$27	31%	$83

Similarly, Figures 4 and 5 show different quantitative impacts on economy due to AVs [21, 23].

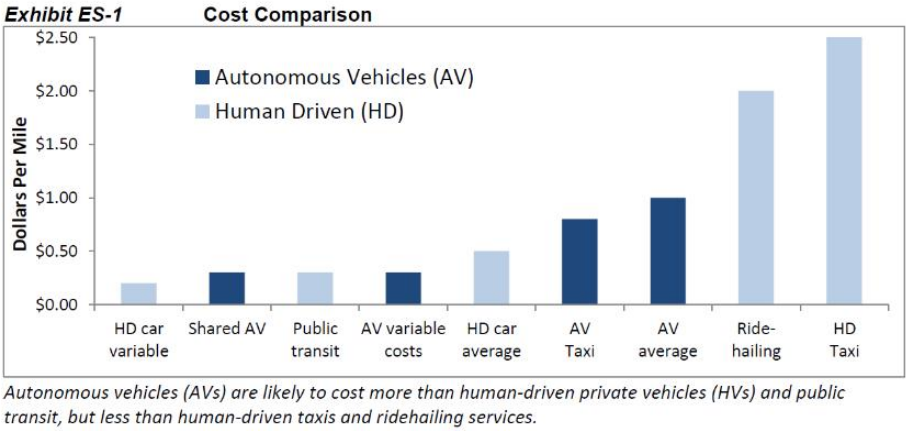


Figure 4: Cost comparison between AVs and Human driven vehicles (Dollars per mile) [22]

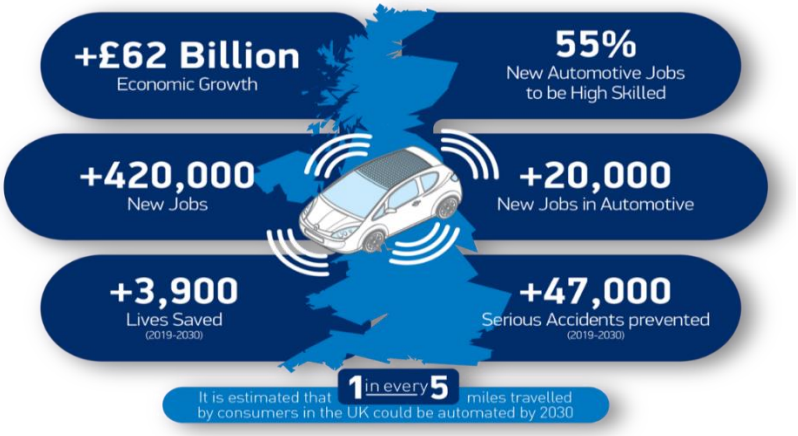


Figure 5: Different quantitative impacts on economy due to AVs [23]

The next section describes the safety/security impacts of this technology.

3. Safety and security impacts

3.1. Recent smart car accidents: In May 2016, a driver was killed while driving a Tesla Model S smart car on the autopilot system [24]. A truck was crossing the highway and the car crashed into the side of the truck. The investigation revealed that the Tesla autopilot system did not recognize the white side of the truck against the bright sky. It was also found that there was no defect in the design or performance of the Tesla autopilot system, but the sensor did not work under the bright sky. This accident highlighted that a constant attention of the driver is required, and the driver must be in a position to take control of the car in emergency situations. Similarly, a woman was killed by an Uber smart car in the street of Arizona [25]. This was the first accident between a smart car and a pedestrian in the US. The driver was inside the car, but the car never generated any alarm for the driver to take control. In other words, smart car failed to detect an object or moving object, or failed to distinguish between the stationary and the moving objects. Similarly, the Tesla smart car got another fatal crash when it failed to detect the stationary object and crashed into a concrete lane divider and burst into flames. The driver died in the hospital [26]. These accidents highlighted that these smart cars are still not ready for a safer drive under all weather conditions, so safer and more secure algorithms are needed.

3.2. Passengers and non-passengers safety: Road crashes cause nearly 1.3 million lives every year, along with 20 to 50 million injuries/disabilities. Since 2014, AVs have 34 reported accidents on the California roads. Auto experts believe that there will be at least 500 million smart cars on the roads till 2025. International study indicates that human error accounts for more than 90 percents of the crashes, including 10 to 30 percent user distractions or inattention.

This does not mean that drivers are the only cause of accidents, but it indicates that human error is one of the main dominant factors in road crashes. AVs have the ability to reduce or eliminate human errors, i.e., enhancing road safety, and minimize accidents. The safety benefits of AV also extend to other road users, i.e., pedestrians and cyclists, since level 4 and 5 AVs have the ability to detect the objects and take necessary safety actions automatically. Therefore, the social acceptance of this technology heavily depends whether AVs are safe or not [22, 27, 28].

3.3. Cyber security: AVs depend on a wide array of electronics, sensors, and software components which are vulnerable to cyber attacks. Cyber security vulnerabilities could not only risk physical safety, but also compromise sensitive personal data. There can be physical threats, software failures, denial of service, eavesdropping, hijacking, malicious code, identity fraud, etc. These threats can lead to smart car compromising with the users needs of security and safety.It is the responsibility of the governments, the industries, the public, and the private sectors to work together to protect the end-user from the cyber attacks. Many such attacks have already happened, i.e., Charlie Miller, and Chris Valasek made a remote attack by taking control of a jeep and sending it off the road and forcing 1.4 million cars to be recalled; the hackers hacked the BMW connected drive car and managed to remotely unlock cars with a recall of 2.2 million cars and attacks on the Volkswagen cars recalling about 100 million vehicles. Similarly, a recent attack on Tesla electric car resulted in software updates for the car operating system [29]. Therefore, the cost of cyber security is becoming a critical issue for the car manufacturers beside reputation damage. These threats have critical impacts on the security, safety and the privacy of the drivers, passengers, and everyone traveling on the road. With the advancement of technology, these attacks are also likely to increase [30]. Robust security measures are necessary to ensure public confidence in automated technology in order to provide confidence that cyber security risks are being controlled or mitigated.

3.4. Legal issues (liability and insurance): The study shows that human mistakes were a definite or probable cause of more than 90 percent of vehicle incidents [31, 32]. The deaths on the road could be reduced by 90 percent by using this technology, as AVs do not drink, do not take drugs, do not use mobile phones while driving, and do not involve any other reckless activities like humans. However, technology is not perfect, and it is still likely that accidents will occasionally occur due to the failure of technology, human driver car interface, maintenance, or other factors, etc. There is a big debate in the case of accidents that who will be liable, i.e., drivers, or cars? Similarly, who will pay for an accident, drivers, or car

manufacturers, especially when there is a system failure? This would ultimately create the ethical dilemma of finding out who is responsible for the accident, and moral decisions might be taken based on the information from the artificial intelligence based algorithms. Similarly, the smart cars are fuel-efficient, but insurance might be high as compared to today's cars especially for the electric version. Insurance rates for electric vehicles can be 21 percent higher as electric cars tend to be more expensive and have expensive repair cost, i.e., trained workers are needed to do work on expensive battery systems. Therefore, the concept of car insurance will change altogether from what we see today. Insurance rates will be decided more with respect to the risk profile of the vehicle than the risk profile of the driver, which is the reverse of what we see today [33].

3.5. Privacy and data: Due to the IoT, smart cars will generate lots of valuable data which will provide new information and services to consumers with a range of applications, e.g., travel information, planning and decision making for infrastructure, traffic management, remote diagnostics, maintenance, continuous improvement of automated driving features, accident investigation, and determining faults, etc. This data will benefit the consumers, the manufacturers, and the government agencies. Some of the data generated by an automated vehicle could be personal data. There are public concerns that who will collect/store this data, and how manufacturers, governments or other corporations will use this consumer generated data. One of the concerns is that the automated vehicles might have an accurate satellite positioning capability, and could also map the location history. There is also a risk that cyber attacks could expose personal data into malicious, or improper use if data reaches in the hand of terrorist organizations, hackers, disgruntled employers, and hostile nations, etc. In order to build the confidence of the public, the personal data must be safe in order to make it compatible with the community expectations [27].

3.6. Enhanced safety measures: Whenever a driver will lose his attention while driving, the risk of accidents will increase, which can be minimized by the use of fatigue detection system of the smart cars. The smart car analyzes the way in which driver uses the smart cars and performs certain functions, i.e., use of the steering wheel, accelerator pedal, brakes, etc. This enables the smart cars to understand that whether the driver is tired, distracted, drunk, and even nervous. Therefore, whenever there will be a change in the behavior of the driver, smart cars will activate an alarm for five seconds. If the driver does not take rest within fifteen minutes following the alarm, the alarm will be activated once again. It will ultimately minimize the risk of accidents [33].

3.7. Level 5 automation: Smart cars with Level 5 automation would eliminate driver stress, distractions, recklessness, and risk taking behavior which causes accidents. However, there will be accidents as smart cars will interact with other cars. The system failures and sudden breakdowns in smart cars would be the major cause of the road accidents. This thing is more critical as we are moving towards the deployment of level 5 automation. However, the safety of higher levels AVs on the road remains untested on a large scale, and may not be available immediately.

Table 7 shows a high level summary (qualitative) of the safety/security impacts

Table 7: High level summary (qualitative) of the security/safety impacts

Impact	Scale	Importance	Outcome	Advantage/Disadvantages
Safety/ security	Big	High	Ambiguous	Human error is reduced but still the abilities of the AVs for perfect driving is still low
Privacy	Medium	High	Privacy level will be reduced	More comfort but less privacy can be a good point for some people and bad for others
Legal issues	Medium to high	High	Complicated legal issues will arise & companies will be more responsible (product liability, etc)	Legal issues between people won't be only personal like before

Table 8 shows a high level summary of the safety/security impacts (quantitative) due to the AVs [17].

Table 8: High level summary (quantitative) of the security/safety impacts

Factors	Scale
Total Crashes per year in U.S.	5.5 million
% human cause as primary factor	93%
Economic Costs of U.S. Crashes	$300 billion
% of U.S. GDP	2%
Total Fatal &Incurious Crashes per Year in U.S.	2.22 million
Fatal Crashes per Year in U.S.	32,367
% of fatal crashes involving alcohol	31%
% involving speeding	30%
% involving distracted driver	21%
% involving failure to keep in proper lane	14%
% involving failure to yield right-of-way	11%
% involving wet road surface	11%
% involving erratic vehicle operation	9%

Similarly, Figures 6 and 7 show different quantitative impacts on safety/security due to the AVs [34, 35].

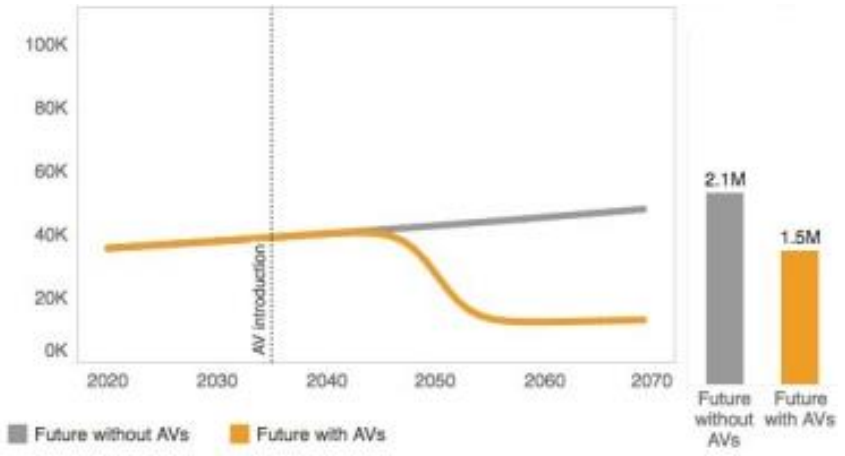


Figure 6: Number of fatalities comparison [34]

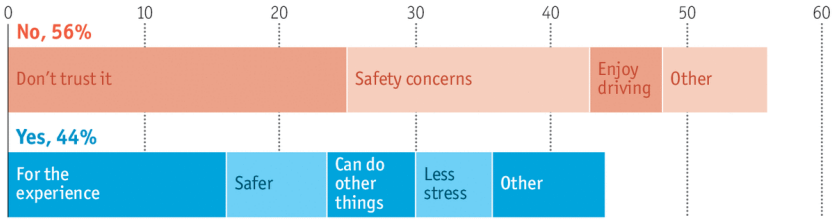


Figure 7: Trust on AV (a survey) [35]

The next section describes the social impacts of this technology.

4. Social impacts

4.1. Impact on mobility: Autonomous vehicles will have a huge impact on the mobility due to the comfort, convenience, and possibility of doing other activities while traveling. This will increase the number of trips which will turn into overall traffic congestion. Moreover, whenever there is an accident on the road, the flow of the traffic is disrupted, but AVs are expected to reduce the accidents up to 90 percent. With fewer accidents, it is also expected that the flow of traffic would increase. Similarly, the new developed infrastructure will provide more facility to the smart cars to travel easily and it will also increase the traffic. The public transport is also expected to vanish, or reduce as one bus will be replaced by many cars. The totality of many cars will also add to the flow of traffic. On the other side, an increase in the traffic might also cause the congestion and lower travel speed on the roads. However, it is still not clear whether all these things will really increase the traffic flow or will have adverse effects. It heavily depends on the characteristics of the AVs as it is not clear whether AVs will really drive faster as we expect or it will be slower. However, all studies generally agree that smart cars will increase the vehicle kilometer traveled [36].

4.2. Accessibility and commutation: The smart car technology will provide people new opportunities and huge ability to do many new things especially in cities. The technology will provide mobility to new user groups (blind, old, physically disabled, or people without driving licenses, etc.), will avoid fixed schedules, and will offer pick-up services at any location, etc. Therefore, AVs will provide an easy and independent access to those people who are always depending on others for commutation. Another major benefit will be the availability of transportation from the last stop of a public transit facility to people's homes, or work places especially for older and disabled people. This will ultimately increase the demand, but it might also slow down the traffic. However, an increase in the AVs would only be possible with advanced infrastructure which will be a great societal benefit also. The expected advantages due to the accessibility should be carefully examined keeping in view the expected loss due to congestion. Changes in behavior and decision making should also be considered due to increased accessibility [10, 19, 22].

4.3. Impacts on work organization: In today's fast paced life, allowing more flexibility in choosing work timings is another source of the connected society. With this technology work can also be done while traveling. This is particularly good for the business sector, which depends heavily on their employees who have to travel longer distances to reach the workplace. In this way, the number of workplace options for an employee is increased, and longer distances to the workplace are also tolerated. This facility can also urge upon the people to live far away from the urban areas, and also provide a facility for the companies to move out from the center of the cities to the suburban area to reduce their expenses [10, 19, 22].

4.4. Impacts on user profiles: One has the choice of high flexibility with his own non-AV private car, but he has to buy a car, pay insurance, bear garage expenses, parking costs, and fuel, etc. Even after bearing all these expenses, one cannot utilize his time in the car productively. Public transport offers more transparent pricing, i.e., weekly or monthly passes. However, the defined set of routes and fixed bus stop locations are also some of the major disadvantages. Therefore, the choice of a private car or public transport can always be seen as a question of availability. Similarly, an individual trip can be made by a car sharing facility with payment, but above mentioned problems can only be reduced in an urban environment. It is expected that car sharing would not produce great success in remote areas [10, 19, 22].

4.5. Smart assistance: Smart cars will gather all information about every driver. It will identify the driver using smart camera and will adjust the settings, i.e., seating, air conditioning, radio frequency, routes, and mirrors to the driver preferences, etc. The driver can also authorize the vehicle for another individual. Furthermore, the smart assistant will also organize the calendar and timetable, determine the optimized routes and stops according to the current traffic, and so on. In this way, smart cars will minimize the risk of the loss of attention and will reduce the accident risks by storing the data about habits [17].

4.6. Better travel management: Through connected vehicles, drivers will receive advance information about traffic, which will enable people to reschedule, postpone, or cancel their journey, etc. Travelers may decide to avoid busy roads, take alternative routes, or use public transport or reschedule their journey. All these options will make the travelers become more fuel efficient and environmentally friendly. People can also move to the public transport, which will reduce greenhouse gas emissions resulting in a significant reduction of fuel consumption, reduced traffic congestion and improved air quality [37, 38].

4.7. Tax issues: Government revenue would decrease by less collection of money for different penalties, i.e., speedy tickets, towing fees, etc, which happens due to human error. The government may impose additional taxes on auto sector, insurance, operators, manufacturers and aggregators to recover the lost revenue opportunities [48].

4.8. Land use patterns: Due to the car sharing and a decrease in the travel cost may reduce the population burden on the cities. People who can afford to buy AVs would prefer to live in suburbs, which may also increase their commutation. However, whether the urban density would increase or decrease would be realized exactly over the longer period of time. [17]

4.9. New driving experience and Public transport reduction: Smart cars will have the ability to interact with the drivers to understand their feelings, and analyze how they are behaving based on some patterns, facial expressions, way of braking, acceleration, drifting, and even the type of music the driver is listening to, and let the car act based on these inputs. Therefore, a smart car will be a secretary for the driver, arranging his appointments, listing his groceries, and driving the car on his behalf while he is not willing. The smart cars will also provide health supervision and safe drive. These things will not only increase the comfort, but will also affect the quality of the lifestyle and the driving experience will be different. Similarly, if huge investment is done in this sector, it can also have a negative effect on the other transport services especially the public transport. AVs may replace the public transport, and it would have an immediate effect on the people with low income, and having no personal transport [39].

4.10. Trust: AVs cannot get a widespread adoption until unless it is assured a trustworthy technology. As many complex systems are involved in the operation of an AV, i.e., GPS, map data, external devices, other vehicles information other than sensors, hardware, software, etc. The question is that how much trustworthy are these different systems?

Table 9 shows a high level summary (qualitative) of the social impacts

Table 9: High level summary of social impacts

Impact	Scale	Importance	Outcome	Advantage/Disadvantages
Mobility	Big	High	Mobility will increase	More comfort, convenience, and possibility of doing other activities while traveling.
Work organization	Medium	Medium	Good for the business sector	Work can also be done while traveling. More choices for time investment
Better travel management	Medium	Medium	Travel management will improve	Will enable drivers to reschedule, postpone, or cancel their journey, etc., based on advanced information's
Land usage	Medium	Small	Ambiguous	Due to the car sharing, and a decrease in the travel cost may reduce the population burden on the cities.

Table 10 shows a high level summary of the social impacts (quantitative) due to the AVs [16].

Table 10: High level summary (quantitative) of social impacts

Aspect	Decrease	Little decrease	No change	Little increase	Increase	Uncertain
Employment (overall)	0.6%	6.4%	52.9%	26.8%	1.3%	12.1%
Employment (transportation)	7.1%	29.5%	30.1%	24.4%	0.6%	8.3%
Cars on road	3.2%	32.3%	25.9%	22.8%	9.5%	6.3%
Segregation	0.6%	12.8%	53.2%	17.3%	4.5%	11.5%
Transit ridership	5.1%	35.7%	27.4%	18.5%	1.9%	11.5%
Equity (mobility)	2.6%	21.8%	28.2%	32.7%	7.1%	7.7%

Similarly, Figures 8 and 9 show different quantitative impacts on social due to the AVs [40, 41].

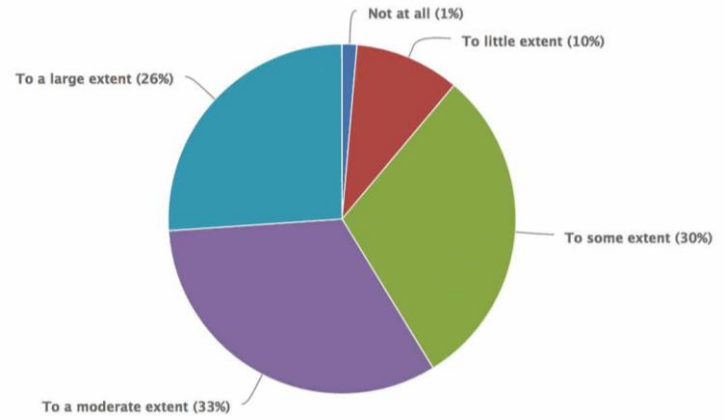


Figure 8: People trust on AVs [40]

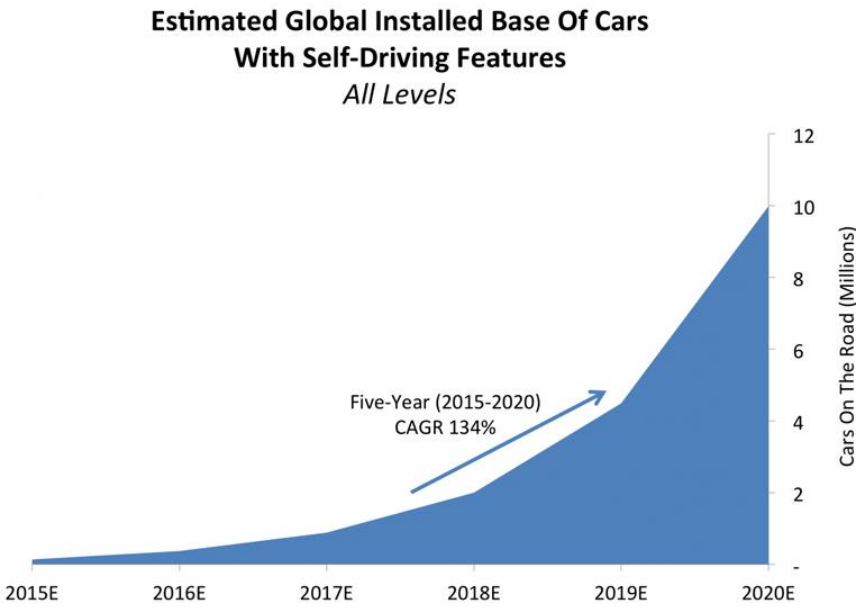


Figure 9: Number of cars with AV features [41]

Previous sections have shown clearly that the autonomous vehicles will have huge impacts on our life and society. There are still many areas in which the outcome of the AVs is also ambiguous. The next section discusses about the challenges and issuesof the AVs.

5. Challenges/barriers of the AVs

5.1. Hardware technology (current developing pace and issues): Hardware technology will determine how quickly the smart cars will be on the road especially level 5 AVs. The innovations in this technology will deliver the required computational power, i.e., graphics and central processing units (GPUs/CPUs). In fact, the hardware technology is already approaching the levels required for well optimized AV software to run smoothly. With the current advancements, the vehicles that adhere up to level 4 automation will probably hit the market in next five years. Sensors have the required range, resolution, and field of vision, but still have significant limitations in bad weather conditions. Radar technology is ready, and has the capability of detection in rough weather and different road conditions. Similarly, the LIDAR system is offering the best field of vision at high levels of granularity. Currently, these equipments are expensive, but inexpensive devices will be commercially available soon after the technology improves. Several high tech companies are claiming to have developed less expensive LIDAR, i.e., within 500 US dollars, but companies still need to understand the optimal number of sensors required for commercial purposes especially for level 5 automation. The technology has been tested in certain feasible conditions, but verifying it under every kind of situation/condition (unmapped areas, or with significant infrastructure/environmental

features, etc.) might take years to ensure complete security and safety. Tuning the system, especially for level 4 and 5 automation to operate successfully in any given situation needs more time and technical expertise [3, 42]. Finally, which technology will be used in the automated vehicles for perception, localization, and mapping is also one of the critical aspects. One approach uses radars, sonars, and camera systems to observe the environment. These devices require less processing power, but they do not assess the environment on a deep granular level. The second approach uses LIDAR system with radar and camera and it is more robust in heavy traffic situations, i.e., provides granular, and high definition (HD) maps, but needs more data processing and computing power. Therefore, experts believe that LIDAR technology will be the technology of the future AVs [44].

5.2. Software technology: One of the major issues is the development of the reliable software. This is also very important as the normal logic, i.e., IF Then statements are not applicable to smart cars as it will take a lot of time and computational resources to handle millions of conditions (data) that the smart car faces. Machine learning needs lots of data to perform different functions (also shows high performance when a lot of data is available), and smart cars (IoT) deal with huge amounts of data. Therefore, one of the best applications to use machine learning is the smart cars which use huge amounts of data. In order to have a secure, safe, and reliable driving experience, and build a high level of trust in the AV users, the industry must develop and employ highly efficient artificial intelligence based algorithms [45].

5.3. 5G: 5G technologies are expected to highly improve the capacity and data speeds of the wireless networks. It will also enhance and develop the technology of the AVs which needs a high speed network to acquire and process the huge amounts of data. 5G networks have already good coverage in urban areas, but recent studies show that dead zones in rural areas may reduce the smart cars to cities only. Therefore, if there is less investment in the rural areas, than AVs will not be able to travel through highways and in rural areas. This situation will be more prominent, i.e., around 2025, when the number of smart cars will increase in rural areas. Therefore, telecom sector would need to ensure that they can provide reliable connections between the most trafficked highways and rural areas [46].

5.4. Power consumption issue: Power consumption is also a big issue with smart cars. A car with just cameras and radar generates 6 gigabytes of data for every 30 seconds. This data will go higher with additional devices, i.e., LIDAR, cameras, etc. For higher level automation, all available data needs to be processed which will ultimately consume huge computing power, i.e., around 2,500 watts which is enough to light 40 incandescent light bulbs. For lower level

automation, on board power requirement increases from 2.8 to 4.0 percent [47]. A smooth and efficient traffic flow will only be possible when autonomous vehicles do not need to recharge themselves after a short interval of time, especially for logistic purpose and highway travel. Therefore, devices should be able to perform data processing quickly and efficiently with less power consumption. Smart cars driving systems provide Eco driving facility, but the computers and sensors can consume enough electricity to negate this green benefit also.

5.5. Regulations: There are civil, criminal, moral, technical, and operator liabilities associated while using AVs. If two self-driving cars crash into each other, then who is liable? Is it the operator sitting idle inside the car or a passenger? Or the manufacturer or supplier or network provider is liable? Similarly, does a special driving license needed for autonomous driving? Does AV have to follow all traffic rules? Or different rules will be needed for AVs? There are a lot of other legal questions around the driverless cars along with these questions, but there are no clear answers. Therefore, lot of legal and legislative work still needs to be done in order to address liability, data ownership and other novel AV problems [43].

5.6. Cost: The hardware technology used for AVs, i.e., sensors, radar, camera, etc., is not very cheap at the moment. Similarly, other high tech experts who are involved in the development of this technology, i.e., software programmer, are also expensive. The auto companies will only be able to sell the AVs at mass level, if it is cheap and affordable to the public, especially to those who need it, i.e., older people, etc. Otherwise it can result into a situation creating a socioeconomic inequality on the roads. Similarly, If the cost drops too slowly, the benefits of the AVs will also not have long lasting effects. Auto analysts believe that the cost premium of an AV should drop by at least around $10,000 to become commercially viable to the end users. At this price, AVs would also be able to make 10 percent market penetration as well, but it may take up to ten years[44].

5.7. Infrastructure: The driverless cars are just only a part of the big picture. This technology needs an established infrastructure to run on it. The auto industry is in dilemma whether to design the driverless cars for existing infrastructure or ask governments to build infrastructure first and then build cars accordingly. Current infrastructure can only handle driverless cars on few designed roadways or changes have to be made, i.e., placing sensors on lane dividers or creating special driverless car lanes, etc [44].

5.8. Ethical problems: Who should be held responsible for the accidents from a moral stand point of view is another aspect of autonomous cars. AVs must make good decisions even in the extreme emergency situations. Is this always possible? It is worth to notice that the engineering problem is substantially different from the hypothetical ethical dilemma. An engineering problem can always differentiate between better and worse solutions, while an ethical dilemma is an idealized constructed state that has no good solution. Therefore, driverless cars need to be designed to be able to differentiate between life and death without human intervention. What a car should do if it has no choice but to

decide between killing a rider or a pedestrian? Currently, this kind of situation is rare, but, as the number of driverless cars will increase, these issues will occur on regular basis [17, 43].

5.9. Public acceptance: Public attitude towards AVs are still in the formative stage. It is not yet clear that how much public desire for the AVs. However, the acceptance of the public is directly related to the safety/security, trust and transparency, privacy, better infrastructure, developed regulations, and the cost of the AVs. It is very difficult that a technology is adopted in a widespread manner, especially if it is not safe and secure to use along with other issues. In one of the survey conducted by the Pew Research Center and Smithsonian magazine in USA found that only 48 percent of American would ride on AVs, if they get a chance [17].

5.10. Social challenges: The technology will really effect different occupations and job markets, i.e., taxi/truck drivers, insurance agents, car mechanics, low skilled labor, etc. Similarly, there will be a change in different business, i.e., car manufacturing, etc. Therefore, the government and other organizations must be ready for this kind of situation.

In the next section, this research proposes recommendations in order to address the current and future challenges faced by this technology as mentioned in previous section.

6. **Proposed policy recommendations**

6.1. Funding required at government levels: Car manufacturers and others have already invested heavily for the research and development of this technology. But, still there are many areas in which the effects of the AVs are ambiguous. Therefore, now it is the time for the governments, transport research agencies, planning organizations, and other stake holders to contribute/fund for the research and development of this technology, so that maximum benefits can be taken from this technology [12].

6.2. Regulations required for liability and security/safety: Liability (civil, criminal, moral, technical, and operator liabilities, etc),and safety/security are the substantial barriers towards the wide spread acceptance of this technology. The manufactures and investors will start pursuing this technology more aggressively, if clear regulations regarding these issues are available as early as possible. Liability regulations should be able to assign responsibility in case of accidents. Countries around the world should work together to deal with these legal issues.At the same time, policy makers should also weigh the effects of extra regulatory actions on AVs, which may be harmful to technological advancement[12].

6.3. AVs certification required: There should be a standardized AVs certification process at least at the government level, and efforts should be made to enhance it to the world wide level. Although, the countries like America has already developed broad principles for

AV testing. But, these testing are currently at the state level only. But, there should be a single document for the adoption by all the states, with the possibility of modifications to suit specific local needs. With such standardized rules, AV manufacturers will be better able to meet national and even international requirements. Moreover, such AV certification will also likelyhelp limit AV product liability [12].

6.4. Understanding AVs impacts: There should be contributions (financial, technical, academic, awareness programs, etc)from organizations, auto industry, local and federal governments, and other stakeholders for the better understanding of this new technology to the public. By creating such awareness, maximized benefits can be obtained from this technology, i.e., climate benefits, energy consumption, land use models, etc [12].

6.5. Managing social effects: The governments and policymakers should have a smooth plan for the coupling of this technology with the society. Therefore, there should be clear and strategic solutions from the governments to handle different social effects of the AVs, i.e., providing alternatives to people losing jobs, taking measures to minimize the negative impacts of AVs on the society, etc [12].

6.6. Developing end user trust: The governments/industry should come up with such standards/regulations to provide high level of transparency and make sure that the privacy of the end user will not be sacrificed at any cost.

6.7. Quality assurance process/programs: New quality assurance processes should be developed in order to monitor the quality of smart cars at all level. The quality of components must be checked from hardware to software and during the process of design and development, ethical aspects must be included. How the decisions are made during critical condition must be given a high importance, i.e., quality of decision making, rules following a certain guidelines, ethical principles, etc. Similarly, how the quality will be maintained during life time must be assured for safety and reliability [17].

6.8. Stakeholders and public interests: In order to take care the interests of all stakeholders and public (all groups, i.e., young, old, working disable, etc), their opinion must be taken into account in the process of design and developments. This will make sure that how much freedom can be provided in decision making process, especially in case of critical conditions. This will determine that how much control a human has in context of the self-driving car. Moreover, the dialogue between automakers and public will decide the priorities and choices of future smart cars [17].

The research has highlighted that the safety/security is the biggest issues in the AVs. The society will not adapt this technology until unless this deficiency is not over. Therefore, in the next section, this research proposes new ideas to for the security/safety of the AVs.

7. Proposed solutions for improvements of safety/security of the AVs

7.1. Regulations for beaconing: It is always difficult to understand the behavior of the surrounding, i.e., vehicles, pedestrian, appearance cues, brake lights, or human behavior (moves right while giving a left signal), etc. We propose an idea of using regular beaconing messages to handle this situation. A regular beaconing should be used at all times for communication (V2V, V2N, V2I, V2X, etc) informing others about the current AV location, speed, heading, etc, along with other messaging. Although, AVs are still using beaconing messages, but, existing vehicles and many newly made vehicles do not have any capacity to send and receive beaconing messages. Therefore, these vehicles do not get any kind of alarm in case of any emergency. We propose a solution to this issue that the governments should make such regulations that all newly made vehicles (whether AV or not) and all existing vehicles must install necessary hardware/software to receive and transmit beaconing messages. The cost of beaconing device is usually less than 50 US dollars [44].

7.2. Fusion based sensor technology: AVs fail to perceive the environment, especially, when there is snowstorm, mirrored reflections, fog, rain, fast moving objects around blind spot, etc. Individual radar, Lidar, sensors, and computer vision system, etc., do not work well to perceive environment under complex conditions. The only safest, easiest, and lowest cost solution is to use the fusion based sensor technology (combination of multiple sensors and deep learning artificial based algorithms) to provide a high level safety in the AVs [47].

7.3. IEEE 802.15.4 MAC protocol: For the MAC layer protocol, IEEE 802.11 and IEEE 1609.4 have been used. We propose the use of another low power, less complex, and low data rate protocol, IEEE 802.15.4 MAC (*added and written in red after modification of the original stack*) for less bandwidth hungry applications for communications, i.e., monitoring and control, remote sensing, etc., as compared to those applications which require a high data rate protocol such as IEEE 802.11. In this way, many low data rate applications can be realized efficiently to further improve the safety by deploying IEEE 802.15.4, which has not been exploited so far for the AVs.

Keeping in view of the challenges, the next section proposes a comprehensive layered architecture for the AVs.

Figure 10 shows the protocol stack for AVs [50].

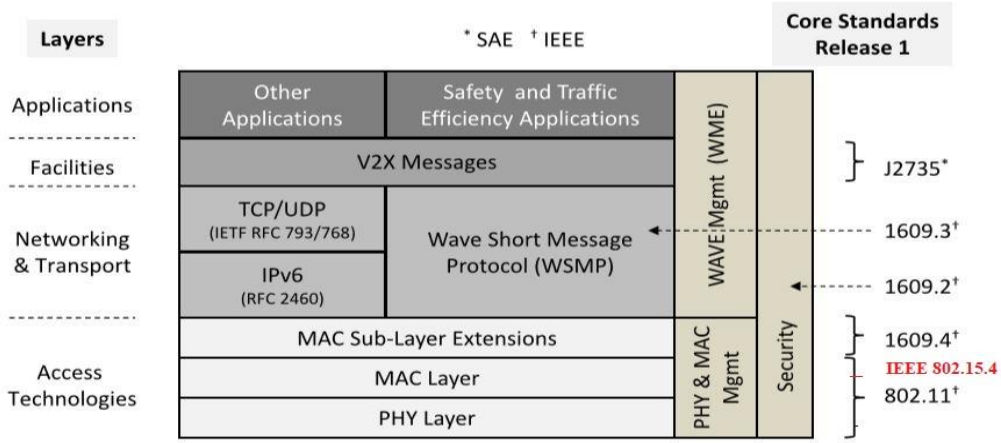


Figure 10. Protocol stacks and core standards for Dedicated Short Range Communication (DSRC) [50].

Proposed comprehensive layered architecture for AVs

Figure 11 shows a proposed comprehensive layered architecture for the AVs [51].

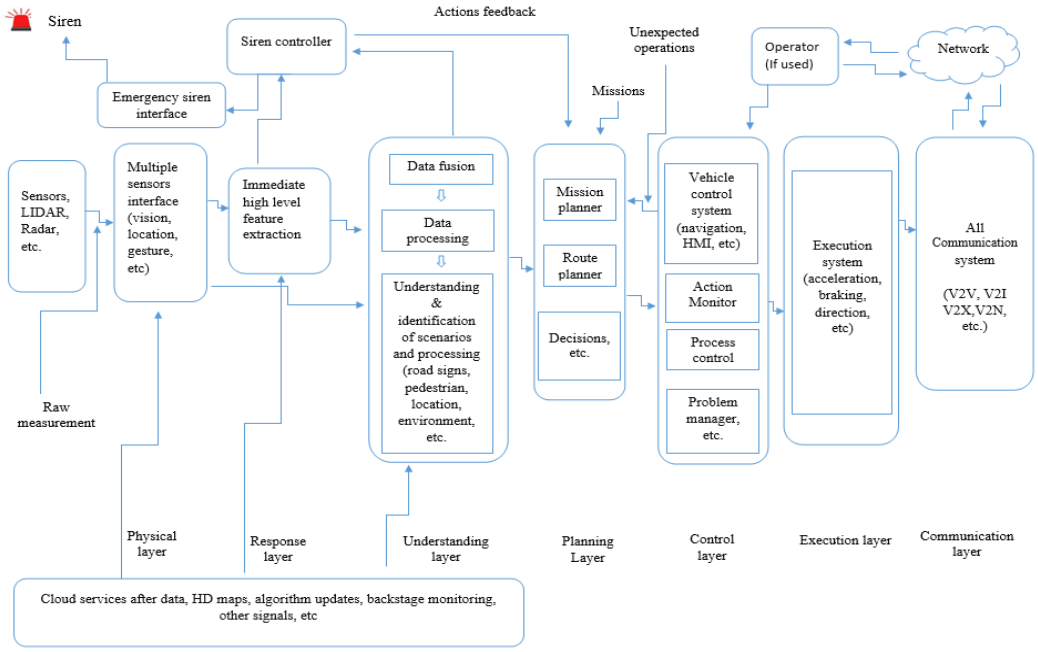


Figure 11. Comprehensive layered architecture for the AVs [51]

According to the architecture, different components can be classified as:

(I) **Physical layer:** This layer will provide a hardware interface with multiple kinds of sensors, i.e., vision sensor, location sensor, gesture sensor, radar, Lidar, and siren interface, etc.

(II) **Response layer:** High level feature extraction will be done at this layer for any emergency action. Deep learning algorithms must be applied at this layer.

(III) **Understanding layer:** Raw data from multiple sensors, and extracted features from the response layer will be fused and processed in this layer to understand the environment, i.e., road signs, pedestrian, location, terrain, environment, etc.

(IV) **Planning layer:** All kind of planning will be performed at this layer, i.e., mission, route, decisions, actions, etc.

(V) **Control layer:** Components in this layer will supervise and control all other integrated systems so that the AV is actually making progress according to the plan and will react in case of any unexpected problem, i.e., faults, obstacles, etc.

(VI) **Execution layer:** Based on the decision from the control layer, this layer will execute the desired actions, i.e., acceleration, braking, direction, etc.

(VII) **Communication layer:** This layer will be responsible for maintaining all kind of communications, i.e., V2V, V2N, V2X, etc.

(VIII) **Similarly,** there can be an operator whether sitting inside the car or outside to control the AVs.

In this proposed architecture, along with fusion based approach, an additional measure has also been proposed for the extra safety/security of the AVs at the response layer. The siren controller at the response layer will receive feedback from the extracted features, and understanding layer, and it will initiate an emergency siren in case of any unusual situation to inform the surrounding vehicles to avoid any accident. This is especially very important in the situations where the time difference between getting watchful of the situation and taking control of the vehicle becomes fatal at certain occasions.

9. Conclusion

This research presents a review of the key impacts of the autonomous vehicle on the society, the environment, the security, and the economy along with recommendations, solutions and a layered architecture. In the first part, the research briefly discusses IoTs, i.e., how IoTs and 5G have enabled the industry to manufacture AVs, and then talks about the working, level of automation, and expectations from the AVs. Secondly, the major impacts of the AVs on the economy, the society, the environment, and the safety are described in detail. The study shows clearly that this technology will have a deep effect on different aspects of our lives, i.e., environment (city planning, gas emissions, fuel consumption, etc.), economy (car industry, logistics, revenue collection, etc.), safety/security (recent accidents, cyber security issues, privacy and data, etc.), and society (accessibility, transportation, jobs, etc.) along with ethical dilemma issues. The study shows that although the technology will have many advantages, there are still definitely many issues with respect to socioeconomic effects. Similarly, in many areas, the impact of AVs is ambiguous. In the final part of this research, current and future challenges/issues of the AVs have been described. Keeping in view that the security/safety is one of the biggest challenges; the research also proposes new recommendations and solutions in order to handle this challenge. The proposed recommendations and solutions provide a comprehensive approach towards the safety/security. With these solutions, the society will be able to adapt the AVs much quickly with full confidence. At the end, the research also proposes a new comprehensive layered architecture for the AVS. The proposed layered architecture provides further special features to provide high level safety and security of the AVs which makes it more safe and reliable as compared to the existing architecture.

Human life is very important, and the study clearly shows that this technology needs a lot of improvements in different aspects, particularly safety/security before humans can have full trust in the AVs. These cars are still not able to predict human movements accurately; the sensors do not work under all conditions, i.e., bright sky (Tesla white truck accident), bad weather conditions, (Uber cyclist night crash), and handling unexpected situations, etc. In order to handle these challenges, the research provides solutions through proposed recommendations, solutions and layered architecture.

The study also highlights the issues with respect to liability and legal challenges to the government and insurance companies. Liability and legal challenges should be defined with respect to both low and high level automation. Similarly, safer and more secure algorithms are needed before these cars can become fully safe to run on the road. Therefore, the research

proposes the safest, easiest, and lowest cost solution by using fusion based sensor technology (combination of multiple sensors and deep learning artificial based algorithms) to provide a high level safety in the AVs.

How much this technology will really affect the society is still unclear and uncertain. This is due the fact that the current research is not empirically oriented. Empirical based research will really be able to tell the true impact on the society, the economy, the environment, and the safety, i.e., jobs, social services, emissions, potential shifts in travel behaviors, data management and privacy issues regarding data generation and sharing, gender dimension of autonomous transport, assigning autonomy to artificial intelligence in critical safety situations, and moral dilemmas, etc. Based on the detailed analysis and impacts of the smart cars, it is very clear that until the smart car manufacturers, the vendors, the industry, and the governments do not improve information among themselves, improve exchanges with security researchers, clarify liability among industry actors, and other legal and environmental issues, it will be very difficult for the rapid widespread adaptability of this technology in the society.

10. References

[1] Alem C, Mesud H. Internet of Things (IoT): A review of enabling technologies, challenges, and open research issues. Computer Networks 2018; 144: 17-39.

[2] Rappaport S, Shu S, Rimma M, Hang Z, Yaniv A, Kevin W, Jocelyn S, Mathew S, Felix G. Millimeter Wave Mobile Communications for 5 G Cellular: It Will Work. IEEE Access 2013; 1: 335-349.

[3] Urooj S, Irum F, Nadeem A. Systematic literature review on user interfaces of autonomous cars: Liabilities and responsibilities. In: IEEE 2018 International Conference on Advancements in Computational Sciences (ICACS); 19-21 Feb 2018; Lahore, Pakistan: pp. 1-10.

[4] Kyle Gershbain. The Future of automotive: Looking ahead to 2025. Concord, Ontario: St. Joseph Communications (SJC), 2017.

[5] Krasniqi X, Hajrizi E. Use of IoT Technology to Drive the Automotive Industry from Connected to Full Autonomous Vehicles. Elsevier 2016; 49: 269-274.

[6] Thomas A , Sick N, Sonja H. On sceptics and enthusiasts: What are the expectations towards self-driving cars.22 Elsevier Transport Policy 2018; 66: 49-55.23

[7] Dorairaj V. The Connected Car. Architecture, Challenges, and Way Forward. New York, US: Sasken Technologies Limited, 2017.

[8] Rashid M ,Fakhrul A. Could Future Wireless Communications Be Harmful. Palmerston, New Zealand: Massey University Project Report, 2017.

[9] Morteza T, Austin L, Hannah R, Shen Q, Ming Xu. A Review on Energy, Environmental, and Sustainability Implications of Connected and Automated Vehicles. Environment Science Technology 2018; 52: 11449-11465.

[10] Sebastian H, Kay W. Recent perspectives on the impact of autonomous vehicles. Zurich, Switzerland: Institute for Transport Planning and Systems, Swiss federal institute of technology, 2016.

[11] JeeryG , Susan S. Automated Vehicles, On-Demand Mobility, and Environmental Impacts. Springer Current Sustainable/Renewable Energy Reports 2015; 2: 74-81.
[12] Zia W, Don M, Paul L. Help or hindrance? The travel, energy and carbon impacts of highly automated vehicles. Elsevier Transportation Research Part A: Policy and Practice 2016; 86: 1-18.
[13] Peter B. AV / ZEV Environmental Health Impact Assessment. Victoria, Australia: Aurecon Australia, 2018.
[14] Matthew B, Kanok B, Guoyuan W. The potential role of vehicle automation in reducing traffic related energy and emissions. In: IEEE 2013 International Conference on Connected Vehicles and Expo (ICCVE); 26 Dec 2013; Las Vegas, NV, USA: pp. 604-608.
[15] Raphael B, Climate and energy impacts of automated vehicles, Goldman School of Public Policy, University of California, Berkeley, June 2014.
[16] Anne Hudson, Jinhua Zhao, Are cities prepared for Autonomous Vehicles ? Planning for technological change, U.S. Local governments Journal of the American Planning Association, May 2019
[17] Daniel J. Fagnant A, Kara K. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Elsevier Transportation Research Part A: Policy and Practice 2015; 77: 167-181.
[18] Daniel J, Kara K. The travel and environmental implications of shared autonomous vehicles, using agent based1 model scenarios . Transportation Research Part C 2015; 40: 1-13.
[19] Clark B, Parkhurst G, Ricci M. Understanding the Socioeconomic Adoption Scenarios for Autonomous Vehicles: A3 Literature Review. Bristol, UK: University of the West of England Project Report, 2016.
[20] Todd L. Are Vehicle Travel Reduction Targets Justied? Evaluating Mobility Management Policy Such As Targets To Reduce VMT And Increase Use Of Alternative Modes. Victoria, Canada: Victoria Transport Institute, 2013.
[21] Lewis M. Clements Kara. Kockelman, Economic effects of automated vehicles, Transportation research record, Journal of transportation board, Jan 2017.
[22] Todd L, Autonomous Vehicle Implementation Predictions: Implications for Transport Planning, Victoria, Canada: Victoria Transport Policy Institute, 2013.
[23] https://www.smmt.co.uk/reports/light-commercial-vehicles-delivering-for-the-uk-economy/, 2019 (accessed 13 September 2020).
[24] Jack S. We need rules for self driving cars. Issues in Science and Technology 2018; 34: 52-57.
[25] Liechtung J. The race is on! Regulating self-driving vehicles before they hit the streets. Brooklyn Journal of Corporate, Financial and Commercial Law 2018; 12: 389-413.
[26] Edmond A, Sydney L, Max K, Sohan D. Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation. MA, USA: Massachusetts Institute of Technology, 2018.
[27] Klauer G, Thomas A. The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data. Washington, USA: National Highway traffic safety administration, 2006.
[28] Dingus L, Jeremy S. Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey. Washington, USA: National Highway Traffic Safety Administration (NHTSA), 2015.
[29] Straub J, Hartman J. Cyber Security considerations for an interconnected self-driving car systems. In: IEEE 2017 System of Systems Engineering Conference; 18-21 June 2017; Waikoloa, HI, USA: pp. 1045-1052.

[30] Orhan M, Sukru O. Review on Cyber Risks Relating to Security Management in Smart Cars. In: IEEE 2018 International Conference on Computer Science and Eng.; 20-23 Sept 2018; Sarajevo, Bosnia: pp. 406-409
[31] Haghi A, Ketabi D, Ghanbari M, Rajabi H. Assessment of Human Errors in Driving Accidents; Analysis of the Causes Based on Aberrant Behaviors. Life Science Journal 2014; 11: 414-420.
[32] Sharon L, Luka M. The Google Car: Driving Toward A Better Future?. Journal of Business Studies 2014; 10: 7-14.
[33] Hevelke A. Responsibility for crashes of autonomous vehicles: An ethical analysis. Springer Science and Engineering Ethics 2015; 21: 619-630.
[34] https://www.wired.com/story/self-driving-cars-rand-report/, 2017 (accessed 12 September 2020).
[35] https://www.economist.com/special-report/2018/03/01/the-success-of-avs-will-depend-on-sensible-regulation, 2018 (accessed 12 September 2020).
[36] Mladenovic N. Engineering social justice into traffic control for self-driving vehicles. Springer Science and Engineering Ethics 2016; 22: 1131-1149.
[37] Scott W, Shogan S. Qiang H. Use of Data from Connected and Automated Vehicles for Travel Demand Modeling. MI, USA: Center for Automotive Research (CAR), 2015.
[38] Thomopoulos N, Givoni M. The autonomous car-a blessing or a curse for the future of low carbon mobility? An exploration of likely vs. desirable outcomes. European Journal of Futures Research 2015; 3: 168-182.
[39] Luis E. Autonomous Vehicles: A Critical Tool to Solve the XXI Century Urban Transportation Grand Challenge. In: 3rd International Conference on Urban Public Transportation; 17-20 Nov 2013; Paris, France: pp. 564-571.
[40] https://bikepgh.org/our-work/advocacy/save/survey/, 2017 (accessed 12 September 2020).
[41] https://self-acquisition2.rssing.com/chan-1415963/all_p2186.html, 2015 (accessed 12 September 2020).
[42] Zhang W. LIDAR-based road and road-edge detection. In: IEEE 2010 Intelligent Vehicles Symposium; 21-24 June 2010; CA, USA: pp. 845-848.
[43] Jamie C. Lidar Just Got Way Better-But It's Still Too Expensive for Your Car. MA, USA: Massachusetts Institute of Technology, 2017.
[44] Victor H. Ultimate Sensor Battle: Lidarvs Radar. Berlin, Germany: Intelligent software engineering, 2018.
[45] Juntae K, GeunY, Youngi K, Kim B. Deep Learning Algorithm using Virtual Environment Data for Self-driving Car. In: IEEE 2019 International Conference on Artificial Intelligence in Information and Communication; 11-13 Feb 2019; Okinawa, Japan: pp. 444-448.
[46] Ian F, Akyildiz S. 5G roadmap: 10 key enabling technologies. Elsevier Computer Networks 2016; 106: 17-48.
[47] James M, Nidhi K, Karlyn D, Paul S. Autonomous Vehicle Technology. California, USA: RAND Corporation, 2016
[48] Raphael B. Climate and energy impacts of automated vehicles. Goldman School of Public Policy, University of California, Berkeley, June 2014.
[49] Preparing for a driverless future. Nshisht Desai Legal and tax consulting associates, September, 2017.
[50] Andreas F. Standards for vehicular communication- from IEEE 802.11p to 5G. Article in e &iElektrotechnik und Information stechnik 2015; 3: 931-949.
[51] Adrian C, Carlos S, Alejandro R, and Pascual C. A Review of Deep Learning Methods and Applications for Unmanned Aerial Vehicles. Journal of sensors 2017; 548-562

Dual-Channel EEG Acquisition Circuit for Vehicular Safety System Based upon Brain-Computer Interface (BCI)

Al-Canaan Amer,
Faculty of engineering, Islamic University of Madinah, KSA
AmerC@iu.edu.sa
and
Uzair Muhammad,
Faculty of engineering, Islamic University of Madinah, KSA
MUzair@iu.edu.sa

Abstract: Electroencephalography (EEG) is the basis for many Brain Computer Interface (BCI) applications, where brain waves are captured, filtered, converted to digital form and analysed. In this paper we report on our design and implementation of wide-band dual-channel EEG acquisition circuit that is suitable for various Brain Computer Interface (BCI) applications. The advantages of the designed system include wide bandwidth covering frequencies from 0.2 up to 171 Hz, dual channel, 2 kHz sampling frequency, Analogue-to-Digital Conversion (ADC), simplicity and low cost. The designed EEG acquisition circuit was designed for BCI vehicular safety system, deep sleep and drowsiness detection applications that rely on analysing the lower frequency bands. However, the EEG acquisition circuit offers a wide bandwidth that makes it possible to analyse gamma brain waves that are related to spiritual and high concentration states of the brain. Compared to other EEG acquisition circuits that are limited to 46 and 100 Hz, our design offers greater bandwidth and resolution for high-frequency brain waves, which enables precise analysis of gamma signals. The designed circuit includes level shifter, limiter and ADC, which is driven by Field Programmable Gate Arrays (FPGA). The interface module with FPGA has been accomplished using a Finite State Machine (FSM), which drives the ADC chip and reads digital EEG data into FPGA. The sampling frequency is set at 2 kHz, which enables high grade BCI systems to be realised including clinical applications. Through the use of FPGA, the designed EEG acquisition circuit provides flexible and programmable DSP features, fast processing, accurate timing, stable sampling rate, and low-power consumption. Hardware Description Languages (HDL) such as VHDL and Verilog are the means to program and build various DSP modules inside FPGA.

Advanced DSP and control algorithms to be implemented with VHDL/Verilog, simulated and synthesised on FPGA are planned for future work.

Keywords: Brain-Computer Interface (BCI), Electroencephalography (EEG), Brain waves, Field Programmable Gate Arrays (FPGA), driver safety control modules, Digital Signal Processing (DSP).

دائرة اقتناء EEG مزدوجة القناة لنظام سلامة المركبات على أساس واجهة الدماغ والحاسب (BCI)

الملخص: تخطيط كهربية الدماغ (EEG) هو الأساس للعديد من تطبيقات واجهة الدماغ والحاسوب (BCI)، حيث يتم التقاط موجات الدماغ وتصفيتها وتحويلها إلى شكل رقمي وتحليلها. في هذا البحث نقدم تقريرًا عن تصميمنا وتنفيذنا لدائرة اكتساب EEG مزدوجة القناة عريضة النطاق مناسبة لمختلف تطبيقات واجهة الدماغ والحاسوب (BCI). تشمل مزايا النظام المصمم عرض نطاق واسع يغطي الترددات من ٠.٢ إلى ١٧١ هرتز، وقناة مزدوجة، وتردد العينات ٢ كيلو هرتز، والتحويل التناظري إلى الرقمي (ADC)، والبساطة والتكلفة المنخفضة. تم تصميم دائرة الاستحواذ EEG المصممة لنظام سلامة المركبات BCI وتطبيقات اكتشاف النعاس والنوم العميق التي تعتمد على تحليل نطاقات التردد المنخفضة. ومع ذلك، فإن دائرة اكتساب مخطط كهربية الدماغ توفر نطاقًا ترديدًا واسعًا يجعل من الممكن تحليل موجات الدماغ غاما المرتبطة بالحالات الروحية والتركيز العالي للدماغ. بالمقارنة مع دوائر اكتساب EEG الأخرى التي تقتصر على ٤٦ و١٠٠ هرتز، يوفر تصميمنا نطاقًا ترديدًا أكبر ودقة أكبر لموجات الدماغ عالية التردد، والتي تتيح التحليل الدقيق لإشارات غاما. تشتمل الدائرة المصممة على ناقل حركة المستوى والمحدد وADC، والتي يتم تشغيلها بواسطة مصفوفات البوابة القابلة للبرمجة الميدانية (FPGA). تم إنجاز وحدة الواجهة مع FPGA باستخدام آلة الحالة المحدودة (FSM)، التي تحركها شريحة ADC وتقرأ بيانات EEG الرقمية إلى FPGA. يتم ضبط تردد العينات على ٢ كيلو هرتز، مما يتيح تحقيق أنظمة BCI عالية الجودة بما في ذلك التطبيقات السريرية. من خلال استخدام FPGA، توفر دائرة الاستحواذ EEG المصممة ميزات DSP مرنة وقابلة للبرمجة، ومعالجة سريعة، وتوقيت دقيق، ومعدل أخذ عينات ثابت، واستهلاك منخفض للطاقة. لغات وصف الأجهزة (HDL) مثل VHDL وVerilog هي وسائل لبرمجة وبناء وحدات DSP مختلفة داخل FPGA.

DSP المتقدمة وخوارزميات التحكم التي سيتم تنفيذها باستخدام VHDL / Verilog ، يتم محاكاتها وتوليفها على FPGA ، مخطط لها للعمل في المستقبل.

1. Introduction

In this section we discuss the theory behind our designed EEG acquisition circuit. In section 1.1, we discuss brain waves and EEF artefacts. In section 1.2, we present BCI. In section 1.3, we discuss different types of EEG electrodes and their placement. In section 1.4, we present FPGA. In section 2, we present related work. In section 3, we present the details of the analogue front-end of the EEG signal acquisition module, which includes passive filters, Instrumentation Amplifier (In-Amp), active filtering and interfacing to the FPGA module. The tests and validation are presented in section 4. The conclusion and future work discussed in section 5.

1.1 Brain Waves

Brain waves are electro-magnetic signals that are commonly sinusoidal and their amplitudes range from 0.5 to 100 μV. Using Electroencephalography (EEG) [1], which enables recording brain waves related to brain activity through non-invasive techniques, with the electrodes placed along the scalp. Although invasive EEG techniques produce higher signal level compared to non-invasive EEG, invasive techniques require surgery and thus are not considered for their lack of versatility and added complexity to BCI systems. The EEG frequency spectrum includes frequency ranges starting from 0.2 Hz up to more than 100 Hz; the brain state of the individual may show certain frequencies more dominant. In this article, we designed and implemented a dual-channel EEG circuit that enables BCI applications that rely on the lower as well as the higher frequencies of the brain waves spectra. Deep sleep and drowsiness detection, for instance, rely on analysing the lower frequency bands, while the higher frequencies are related to high concentration and spiritual states of the brain. However, the circuit offers a wider bandwidth that makes it possible to analyse gamma brain waves that are related to spiritual and high concentration states of the brain. Compared to other EEG acquisition circuits that are limited to 100 Hz, our design offers greater bandwidth for gamma waves, i.e., up to 171 Hz, which enables precise analysis of gamma waves. The upper limit (171 Hz) for the bandwidth was chosen in order to avoid the higher harmonics of 60 Hz above 120 Hz, i.e., filtering only the 60 Hz and its first harmonic 120 Hz is necessary. Brainwaves have been categorised into five basic frequency bands. Studies have shown that certain activities can be correlated with increased in power for specific frequency ranges. For example, drowsiness and fatigue signs can be related to an increase in the θ frequency band from 4 to 8 Hz. An example of extracted θ waves in time domain is shown on Figure 1, which was acquired using our designed and built EEG acquisition circuit.

1.1.1 EEG artefacts

EEG artefacts describe signals, which are of non-cerebral origin. The amplitude of artefacts can be large relative to the size of amplitude of the cortical signals of interest. Some EEG artefacts are useful in various applications including Electrooculography (EOG), which enables detecting and tracking eye-movements. EOG is important in Polysomnography or sleep study, where it is used to diagnose sleep disorders, and is also used in EEG for assessing changes in alertness, drowsiness or sleep.

1.2 Brain-Computer Interface (BCI)

Brain Computer interface (BCI) is a technology that enables analysing brain activity or emotions using EEG and perform relevant actions based upon distinctive brain waves. This enables people with certain mobility issues or injuries to overcome restrictions and enhance their self-support. BCI is also employed in medical diagnosis against sleep disorders, drowsiness and certain brain issues.

1.3 EEG electrodes

EEG signals are picked up using electrodes using silver-plated electrodes. Special gel is usually used to ease electrical contact with the skin. Therefore, electrodes are to provide good electrical contact with skin. Different types of electrodes are available, some of which are disposable.

Reusable and disposable electrodes have been procured. The disposable type was not used since the number of electrodes was only enough for several trials.

Reusable electrodes can be utilised as dry or wet. Comprehensive comparison between dry and wet electrodes in [2] justifies the use of dry electrodes over wet electrodes. The following reusable electrodes have been considered:

1. Ag/AgCl hair penetrative electrodes, which are made of plastic material coated with a thin layer of silver.
2. Silver-plated cup dry electrodes.

After some experimentation, the hair penetrative electrodes have shown some wear signs, where the coating material was scratched off. Thus, we opted for the reusable silver plated cup electrodes (Ag) for their outstanding durability. To increase versatility, we have utilised the dry cup electrodes, i.e., without any gel or any additives.

1.3.1 Placement of electrodes

We have utilised two pairs of electrodes, i.e., two channels to capture brain waves. The number of channels determines the number of electrodes. Each channel requires one differential signal, which requires one pair of electrodes. In our design, we opted for a dual-channel circuit in

order to fulfil design constraints related to versatility and ease of use, which was confirmed using practical implementations of BCI systems. By using two channels only, the designed circuit has produced decent EEG signal quality that enabled the classification modules to achieve high accuracy.

Initial tests were made with dry electrodes [2] fixed using plaster tape on the forehead. Later on, a 3-D printed Ultracortex Mark IV frame available from OpenBCI provides good contact and support for the electrodes. The silver-plated electrodes were easy to fix on the front nodes of the frame. The electrodes of both channels on the frontal lobe are placed according to the 10-20 system as follows:

Channel 1: Differential input signal between Fp1 and Fp2 locations.

Channel 2: Differential input signal between the left ear clip and Fpz.

1.4 Field Programmable Gate Arrays (FPGA)

Field Programmable Gate Arrays (FPGAs) are Integrated Circuits (IC) that incorporate matrices of Configurable Logic Blocks (CLB) connected via programmable interconnects. Design on FPGAs can be reprogrammed or tuned many times to accomplish desired design improvement for. Programming an FPGA is accomplished through special Hardware Description Languages (HDL), such as VHDL [3]. The designed EEG circuit enables interface with FPGA, which has been accomplished using a Finite State Machine (FSM) in order to drive the ADC chip and read digital EEG data into FPGA. This makes the FPGA ready to implement DSP modules that we planned for future work.

2. Related Work

The work in [4] provides a single-channel low-cost EEG circuit for BCI system. Our work can be distinguished from the work in [4] in several aspects, with regard to simplicity and bandwidth, since our design is simpler with less amplification stages, without opto-coupling, without driven-right-leg (DRL) circuit and with a second additional channel.

The authors in [5] proposed a circuit design for extraction of EEG signals. Our work can be distinguished from their work in different ways; firstly, our design uses only one instrumentation amplifier per channel, which makes our design more cost effective. Another difference is that the total gain relatively high (10000-20000). We believe that excessive gain may lead to saturation and distortion due to exposure to noise signals.

In [6], a design of EEG signal acquisition system using Arduino MEGA1280 and EEG analyser is proposed. Our work can be distinguished from their work in different ways; firstly, our

design offers higher bandwidth (171 Hz). Secondly, our design proposes only 2 channels, compared to 16 channels, to increase versatility and ease of use.

The authors in [7], the authors propose an implementation of analogue portable EEG signal extraction system. Our work can be distinguished from the work in [8] in several aspects; firstly, we designed a 4th order active filters for the band-pass filters, which reduced the number of components. Although using higher order filters provides better rejection of unwanted frequencies, we opted for 4th order active filter since it requires only two op-amp stages, i.e., one chip of dual op-amp, which is enough to improve the SNR of the signal. Another reason, is the foreseen subsequent digital filtering on FPGA and software, which is more efficient. Secondly, the cut-off frequency of the system is limited to 100 Hz, whereas our design offers bandwidth covering frequencies between 0.18 and 171 Hz. Thirdly, we utilise a pair of 3.7 V batteries power source instead of dual 9 V, which means lower battery dimensions and weight. In [8], the authors propose a wireless 8-channel digital active-circuit EEG signal acquisition system that uses dry sensors. The total gain of the designed circuit in [8] is 9752, which is adequate and slightly higher than the gain of 9333 in our design. Generally, higher gain may lead to saturation and distortion of EEG signal. The bandwidth is 125 Hz compared to 171 Hz. The power supply is 3 V fed from 3.7 V batteries, while in our design we have dual 3.7 V power supply and 3.3 V for the ADC and digital interface. The analogue filters we have designed have a unity gain to simplify construction and reduce the number of components. Also this makes adjusting the gain at a single stage (the In-Amp, in our case). Another difference, is the number of channels is 2 in our design, while is it 8 in [8]. We believe that using 2 channels is sufficient for many BCI applications and offers more versatility. A final difference is cost, their circuit utilises compact modules, Flexible Printed Circuit Board (FPCB) and expensive In-Amp and op-amps, while we use inexpensive In-Amp and op-amps without FPCB.

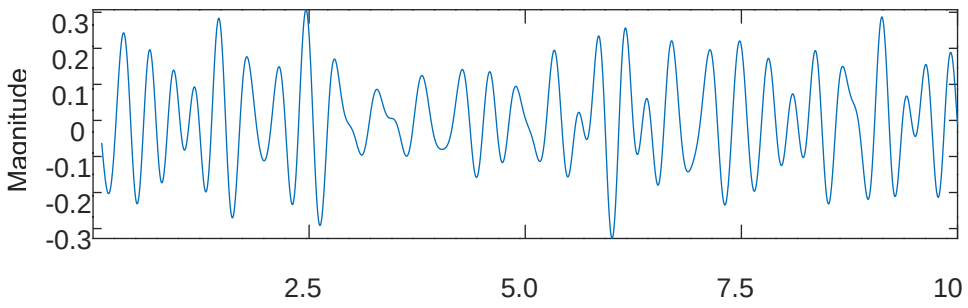


Figure 1: Theta waves (4-8 Hz) captured using the designed and built EEG acquisition circuit.

EEG headsets are available from different manufacturers notably OpenBCI [9] and Emotiv [10], of which the sampling frequencies are 250 and 128, respectively. The Emotiv products are intended only for personal use or research. At their best, the Emotiv cover the frequency range from 0.16 up to 45 Hz. Although the sampling rate is at best 8 kHz, it is down sampled to 128 Hz. Both sampling rates offer approximately the minimal acceptable rate of 2.5 times the highest frequency of interest [11].

On the other hand, since both types of headset are microcontroller-based, the sampling frequency drifts from the nominal value [10] since the hardware does not include a crystal oscillator. In our design, the use of FPGA board to drive the ADC eliminates these shortcomings of precise clock timing based upon a master crystal oscillator and dedicated Digital Clock Modules (DCM). Although most of the available headsets are adequate for research, they lack the precision for the higher frequency content of EEG waves due to the low sampling rate. We believe that the sampling frequency should be much more than double the highest frequency of the targeted signal. Therefore, the sampling frequency used in our design is 2 kHz, for it is more than 10 times the highest frequency component in our EEG system, i.e., 171 Hz. However, higher sampling rates of 2 kHz and even more are actually utilised in costly clinical EEG equipment.

Considering certain design objectives, the designs in [4], [5], [7], [8] and [9] fit into several application domains where they prove to be adequate, for example when the ease of use, cost and quick set-up are not mandatory. This includes BCI medical diagnosis equipment. However, our design objectives include BCI applications where cost, precision, small number of electrodes is preferred. For example, drowsiness detection and smart home control of are two major BCI applications that are targeted by our design objectives.

The number of channels was chosen based upon practical implementations of designs of a drowsiness-detection in vehicular safety system as well as other BCI applications, such as smart-home. In both applications, the dual-channel circuit have produced descent EEG signals that were successfully classified with high precision rates.

3. Analogue front-end of the EEG signal acquisition module

Our designed dual-channel EEG acquisition circuit is shown in Figure 4 is based upon Instrumentation Amplifier (In-Amp) circuit and different types of analogue filters to remove undesired noise. Level shifting and limiter circuit usually follow this circuit in order to convert the captured signal into digital form, i.e., Analogue-to-Digital Conversion (ADC), which enables Digital Signal Processing (DSP) that can be performed using FPGA or microcontroller.

An FPGA module has been successfully connected to the EEG circuit to drive the ADC chip and read EEG data into FPGA, which enables DSP modules to be implemented on FPGA. For the sake of brevity, the FPGA module is not explained in detail in this paper.

The dual-channel EEG acquisition circuit has been assembled on the Electronics Explorer board (available from Digilent), as shown in Figure 5, in order to profit from its integrated test bench for electronics projects. It includes mainly a 4-channel oscilloscope, multi-meter, signal generator, frequency spectrum analyser and different types of filters. The board was used to validate the EEG signals and export digital signals in data file for DSP using Octave, for example. The output of DSP can be viewed in Figure 6, where the attenuation due to the analogue Notch filter can be viewed around the 60 Hz point. The digital filtering effect of 60, 120 and 171 Hz is visible on the same figure with steep edges (clearly visible for the 120 and 171 Hz).

3.1 Passive LPF Since the required gain of the acquisition circuit is relatively high (several thousands), it is convenient to remove part of the undesired noise at the input of the In-Amp using a passive LPF with $f_c = 450\ Hz$ (composed of R1a1, R1b1, C1a1, C1b1 and C1b2, as shown in Figure 3) for both electrodes of channel 1. The value of f_c should be far from the frequencies of interest. For channel 2, the LPF is composed of R1a2, R1b2, C1b3, C1 and C1b4, as shown in

Figure 2. The values of R1b2 and R1a2 are tuned experimentally to produce most adequate results.

3.2 Instrumentation Amplifier (In-Amp) In-Amp is an electric circuit that is usually composed of 3 op-amps, of which two perform buffering with high-input impedance while a third op-amp serves as difference amplifier. The gain of In-Amp is determined by changing resistor R2a1, for channel 1 as shown on Figure 3).

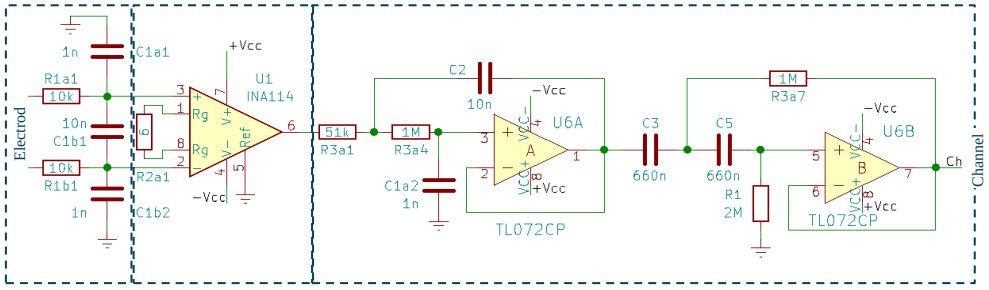


Figure 2: Main block diagram of the dual-channel EEG acquisition circuit.

The total gain of EEG acquisition circuits may be around 10000. We have designed the In-Amp circuit from scratch using two approaches; using discrete op-amps and using dedicated In-Amp chip, as shown in Figure 3. Although, the In-Amp with discrete op-amps produced

coherent EEG signals, high CMR, compactness and gain precision of dedicated In-Amp ICs, such as the INA114 IC, outperform In-Amp with discrete op-amps. Therefore, the INA114 In-Amp has been adopted in the final prototype as shown in Figure 3 as well as in

Figure .

The gain of the INA114 In-Amp for channel 2, as shown in

Figure 4, is determined by R2a3 6-Ω resistor, which produces gain around 9333. The presence of the preceding passive RC filter stage attenuates slightly the input signal.

3.3 Active filtering The raw EEG signals have low amplitude compared to surrounding noise, such as the 60-Hz fundamental frequency and its harmonics. Therefore, we have built double twin-T notch filter stage for channel 2 only, as shown in

Figure . For channel 1, the DSP module utilises Finite Impulse Response (FIR) digital filtering to remove the 60-Hz noise and its first harmonic, i.e., the 120-Hz. To tune the notch filter, we have connected resistors in series to get desired response. Channel 1 was designed such that it can be utilised in 50 or 60-Hz systems, since it does not include notch filters that undertake rejecting the 50 or 60-Hz power line noise.

Although test results showed that both channels offer good results, channel 2 offers better Signal to Noise Ratio (SNR) due to analogue filtering for the 60-Hz noise. Additional digital filtering can be applied to raw data using software or using FPGA in subsequent stages. The output of the oscilloscope can be exported to data file that can be processed using Octave to perform DSP. The frequency response of the EEG signal is shown on Figure 6, where the analogue filtering and digital filtering effects are visible. Obviously, the digital filtering using FIR filters is more efficient than analogue filtering.

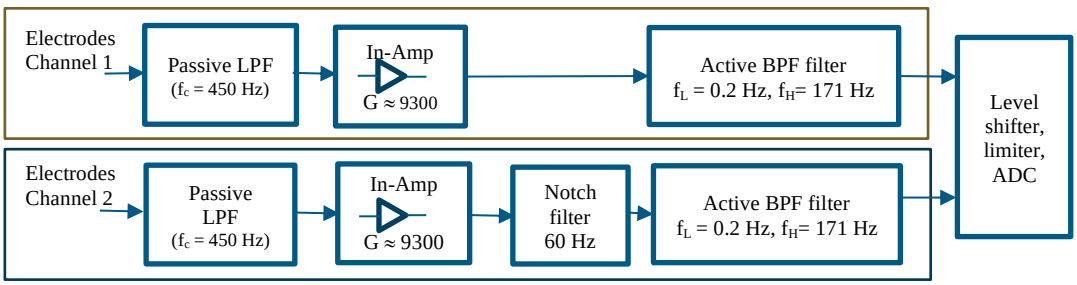


Figure 3: The schematic for channel 1 circuit showing the LPF, In-Amp and BPF.

Both channels contain BPF to further remove undesired frequencies outside of the 0.2-217 Hz frequency band. The design of active filters can be accomplished using online tools [12].

3.4 Interfacing with FPGA digital modules

Before feeding the analogue signals to the ADC, the level of the EEG signal should be shifted and limited within 0-3.3 V. This is accomplished using a level-shifter and limiter circuits to limit the signal below +3.3 V, which is the maximum supply voltage for the used ADC. For the sake of brevity, the schematic of this interfacing circuit is not shown in this context. The ADC enables further DSP processing in FPGA or microcontroller in order to perform additional complex analysis operations, such as Fourier and Wavelet transforms.

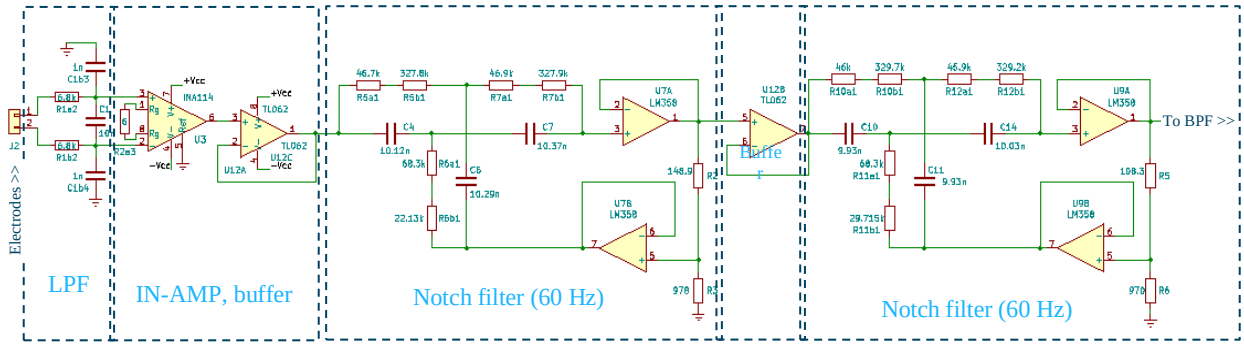


Figure 4: The schematic for channel 2 circuit showing the LPF, In-Amp and 60-Hz Notch filters.

4. Tests and validation results

We have performed several test scenarios for the various sections of the designed circuit. The notch filters are sensitive part to resistors. The test scenarios were performed during the breadboard early phase and after building the prototype on PCB, which is shown on Figure 7.

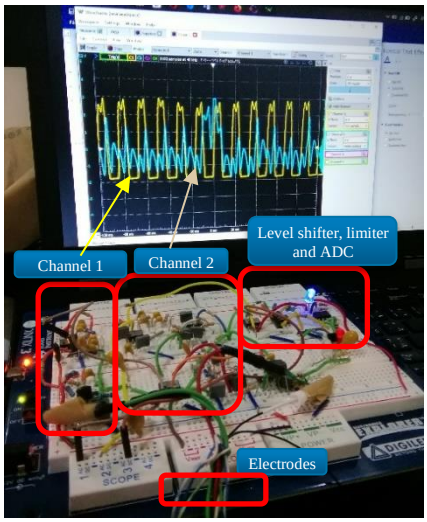


Figure 5: Testing the dual-channel EEG acquisition circuit using oscilloscope during early design phase.

A sample output of captured EEG signals for both input channels is shown on Figure 5, which describes the output of both channels 1 and 2 in time domain, while Figure 8 shows the frequency spectrum of channel 2 as viewed on the oscilloscope, which shows clearly the effect of the analogue filters used; the magnitude of the 60 Hz is less than that for 120 Hz and beyond 120 Hz the effect of LPF is visible. On the same figure, the harmonics of 60 Hz are visible with peaks at 120, 180, 240, 300 and 360. Since the frequency range of interest is below 171 Hz, we have to deal only with 60 and 120 Hz.

After the analogue circuitry was validated, the first PCB prototype was designed and manufactured. Eventually, the EEG board was successfully integrated with BCI applications that take the output data of the EEG board and process it offline in order to perform training of different classification modules. Using the built circuit, a considerable dataset of recordings was established over several months for several persons and different classes related to brain activity. For the sake of brevity, the details of the designed BCI applications are left to future article of ours.

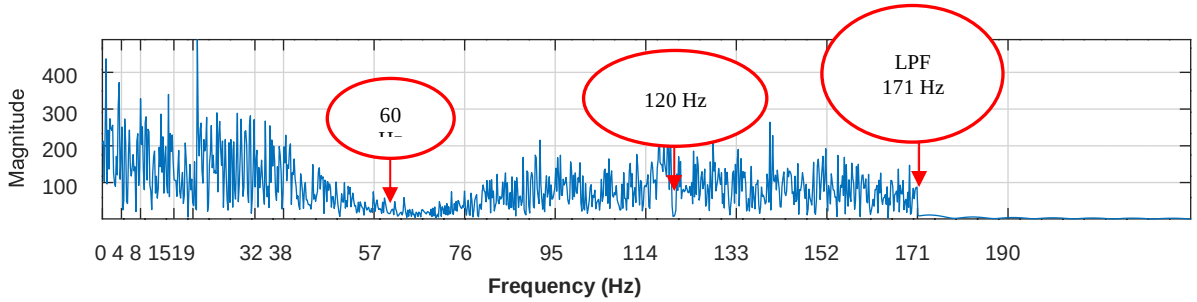


Figure 6: Frequency spectrum showing removal of the 60 and 120 Hz noise and frequencies beyond 171 Hz.

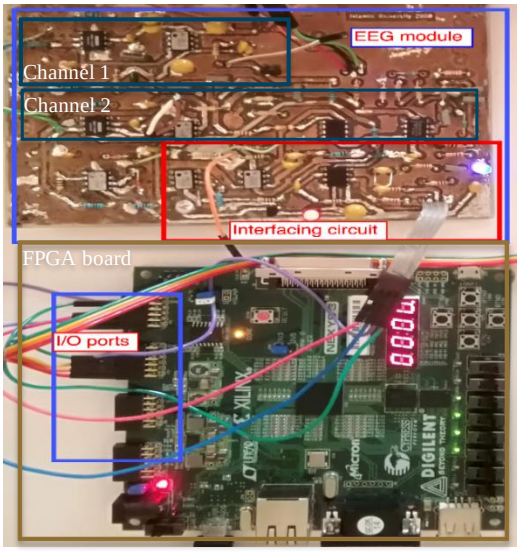


Figure 7: PCB of the designed EEG acquisition circuit connected to FPGA board

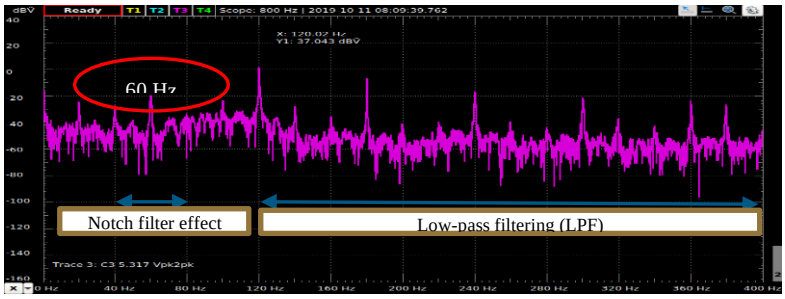


Figure 8: Frequency response, showing the attenuation of the 60 Hz signal and the effect of the BPF.

5. Conclusion

EEG signals are employed in many BCI applications including driver safety, smart-home

Figure 8: Frequency response, showing the attenuation of the 60 Hz signal and the effect of the BPF.

control and medical diagnosis. We have demonstrated our design and implementation of a dual-channel EEG acquisition. Channel 1 does not include notch filters, thus it can be used in 50 and 60-Hz power systems, while channel 2 offers active notch filters to attenuate the 60-Hz noise. The built EEG acquisition prototype is based upon dedicated In-Amp that is followed by BPF and Twin-T notch active filters. The output of the analogue circuit is connected to ADC circuitry that is controlled by FPGA, which enables further DSP modules to be designed and implemented on FPGA. The interface with FPGA kit has been designed, and implemented using FSM.

The advantages of the designed circuit include wide bandwidth covering frequencies from 0.2 up to 171 Hz, dual channel, flexible sampling frequency (2 kHz is recommended), ADC conversion, compliance with 50 and 60 Hz power systems, precise timing and sampling rate, simplicity and low cost.

Introducing programmable DSP features can be accomplished using FPGA that provide flexible and powerful platforms for DSP that employ parallelism and offer low-power consumption. Hardware Description Languages (HDL) such as VHDL and Verilog are the means to build modules inside FPGA. Algorithms of control modules can be implemented with VHDL/Verilog, simulated and synthesised on FPGA.

To fully test the EEG acquisition circuit, it has been manufactured on PCB, integrated with different classifier modules for a drowsiness-detection vehicular safety BCI system as well as smart-home control. Test result were coherent and proved the soundness of design, which are detailed in other articles of ours. The designed EEG acquisition circuit enables the design and development of other BCI systems, such as medical diagnosis for sleep disorders.

All test scenarios were successful, a prototype PCB was built and the interface module with FPGA has been completed and prepared for further DSP modules to be realised.

The future improvements to the circuit include using quad op-amps, integration of complex DSP modules on FPGA and integrate a wireless module to communicate with BCI systems using Wi-Fi or Bluetooth.

6. References

[1] A. Ortiz, «Main features of the EEG amplifier explained,» Bitbrain, April 2020. . https://www.bitbrain.com/blog/eeg-amplifier. (Accessed 23 June 2020).

[2] H. Hinrichs, M. Scholz, A. K. Baum, J. W. Y. Kam, R. T. Knight и H.-J. Heinze, «Comparison between a wireless dry electrode EEG system with a conventional wired wet electrode EEG system for clinical applications,» *Scientific Reports*, т. 10 , № 5218, 2020.

[3] D. Pellerin, «An Introduction to VHDL for Synthesis and Simulation,» . http://www.uco.es/~ff1mumuj/h_intro.htm. (Accessed 23 June 2020).

[4] L. Zhang, X.-j. Guo, X.-p. Wu и B.-y. Zhou, «Low-cost Circuit Design of EEG Signal Acquisition,». *6th International Conference on Biomedical Engineering and Informatics (BMEI 2013)*, 2013.

[5] L. Wenqiang, Q. Lijun, F. Qifu и L. Huan, «The circuit design for extraction of electro encephalogram signals,». *International Forum on Energy, Environment and Sustainable Development (IFEESD 2016)*, 2016.

[6] D. Saptono, B. Wahyudi и B. Irawan, «Design of EEG Signal Acquisition System Using Arduino MEGA1280 and EEGAnalyzer». *MATEC Web of Conferences ICMIE 2016*, 2016.

[7] M. E. Sahin, Y. Ucar и F. Termurtas, «An Implementation of Analog Portable EEG Signal Extraction System,». *2015 9th International Conference on Electrical and Electronics Engineering (ELECO)*, Bursa, 2015.

[8] C.-T. Lin, C.-H. Liu, P.-S. Wang, J.-T. King, L.-D. Liao, «Design and Verification of a Dry Sensor-Based Multi-Channel Digital Active Circuit for Human Brain Electroencephalography Signal Acquisition Systems» *Micrmachines,* т. 10, № 72, 2019.

[9] Strikroot, «How steady is Cyton's sampling rate? [resolved],» OpenBCI, 19 February 2019. . https://openbci.com/forum/index.php?p=/discussion/2329/how-steady-is-cytons-sampling-rate-resolved. (Accessed 22 October 2020).

[10] Emotiv, «Headset Comparison Chart,» Emotiv, 2020 . https://www.emotiv.com/comparison/. (Accessed 22 October 2020).

[11] E. Smith, «EBME & Clinical Engineering Articles-Introduction to EEG,» ebme. https://www.ebme.co.uk/articles/clinical-engineering/introduction-to-eeg. (Accessed 22 October 2020).

[12] A. Devices, «Analog Filter Wizard,» Analog Devices, 2020. . https://tools.analog.com/en/filterwizard/. (Accessed June 2020).

[13] M. O'Sullivan, A. Temko, A. Bocchino, C. O'Mahony, G. Boylan, E. Popovici «Analysis of a Low-Cost EEG Monitoring System and Dry Electrodes toward Clinical Use in the Neonatal ICU,» *Sensors*, June 19(11), № 2637, June, 2019.

Synthesis and Characterization of Nano Au/NPAA Templates

Mohamed Shaban

Nanophononics and Applications Labs, Department of Physics, Faculty of Science, Beni-Suef University, Egypt

Department of Physics, Faculty of Science, Islamic University of Madinah, KSA

mssfadel@aucegypt.edu

and

Sobhi M. Gomha

Department of Chemistry, Faculty of Science, Cairo University, Egypt.

Department of Chemistry, Faculty of Science, Islamic University of Madinah, KSA

s.m.gomha@gmail.com

Abstract: A modified two-step anodization process was applied for preparing highly ordered nanoporous anodic alumina (NPAA) templates. The interpore distance and the diameter of the NPAA pores were, 100 nm and 80 ± 2 nm, respectively. Then, Au sputtering was carried out over the fabricated template using r.f. magnetron sputtering. The morphological characterization showed that hexagonal Au nanoarrays have been deposited on the top surface of the NPAA template. Simultaneously, the interior walls of the NPAA pores have been decorated with fine Au nanoparticles. The Au-coated samples showed brilliant and statured structural colors due to the surface plasmon resonance (SPR) enhancement of NPAA optical interference. The SPR wavelength is increased from 466 nm to 512 nm by increasing the Au deposition time from 1 to 2 min.

Keywords: Nanoporous membranes; Au nanostructures; Surface plasmon resonance; r.f. magnetron sputtering

توليف وتوصيف قوالب Nano Au / NPAA

الملخص: تم تطبيق عملية أنودة معدلة من خطوتين لإعداد قوالب أنوديك ألومينا نانوية (NPAA) عالية الترتيب. حيث كانت المسافة البينية وقطر مسام ال NPAA، ١٠٠ نانومتر و٨٠ ± ٢ نانومتر، على التوالي. بعد ذلك، تم ترسيب جسيمات الذهب النانوية على سطح القالب المصطنع باستخدام طريقة (r.f. sputtering). أظهر التوصيف المورفولوجي أن مصفوفات نانوية سداسية الترتيب من جسيمات الذهب النانوية قد ترسبت على السطح العلوي للقالب. في الوقت نفسه، تم تزيين الجدران الداخلية لمسام ال NPAA بجسيمات الذهب النانوية الدقيقة. أظهرت العينات المطلية بـجسيمات الذهب ألوانًا هيكلية رائعة وثابتة نظرًا لتعزيز رنين البلازمون السطحي SPR مع التداخل البصري لـ NPAA كما لوحظ زيادة الطول الموجي ل SPR من ٤٦٦ نانومتر إلى ٥١٢ نانومتر بزيادة وقت ترسيب الذهب من ١ إلى ٢ دقيقة.

1. Introduction

During the recent decade, great efforts have been directed toward the design and characterization of the nanoporous templates because of their wide range of applications [1-4]. Nanoporous anodic alumina (NPAA) is an interesting porous template with a bright color in the visible light range due to the interference of light [1]. The saturation of the color is very low. To get a highly saturated color and to extend the applications of NPAA, plasmonic nanostructures were applied to coat the nanopores of the NPAA [1, 5].

Gold (Au) is one of the most important plasmonics nanomaterials. The design and optical characterization of Au nanostructures have been extensively addressed [6-9]. The existence of the localized surface plasmon resonances is considered one of the most important features that make Au nanostructures a key material in many applications. The positions of the localized surface resonances are size- and shape-dependent. Also, the surface plasmon resonance coupling shifts the resonance wavelength and splits the degenerated modes for regular Au nanoarrays. So Au nanoarrays can be applied for guiding, enhancing, emitting, and modifying the optical response for different applications including nanoplasmonic-photonic crystals[10], surface-enhanced Raman scattering (SERS)-based sensors [11], and nanoctylasts[12]. Consequently, the controllable deposition of Au nanoarrays on NPAA templates to produce Au/NPAA nanoporous structures is required to address the coupling between surface plasmon of Au nanoarrays and NPAA interference.

To deposit Au nanoarrays on the top surface of NPAA, a deposition technique such as radio frequency (r.f.) magnetron sputtering was applied because it permits the accurate controller of size, shape, and the composition. But, Au nanoparticles are also deposited on the interior walls of the pores of the NPAA template during the deposition process. Therefore, the full optical characterization of the Au/NPAA nanoporous arrays is complicated because of the existence of more than one Au nanoarray, Al substrate, and the electromagnetic coupling among the closest Au nanoparticles. Based on the previous reasons it is imperative to study the design of Au/NPAA nanoarrays experimentally and to correlate between the SPR positions and the morphological parameters of Au nanoarrays.

In this article, a modified two-step anodization process was applied to design NPAA templates. Then, the simultaneous decoration of the interior wall of NPAA nanopores and the growth of Au nanoarrays on the top surface of the NPAA template was performed by r.f. magnetron sputtering. Then, the morphological properties of NPAA and Au/NPAA are studied using field emission scanning electron microscope (SEM). Also, the optical properties are

studied using UV/Vis/NIR spectrophotometer. Moreover, the role of localized and propagating surface plasmons (LSP and PSP) on the enhancement of optical interference between the light beams reflected from the top and bottom surface of Au/PAA film was discussed.

2. Experiments: Fabrication and Characterization

Al foil with purity 99.99% has been used as a starting material for the fabrication of nanoporous anodic alumina (NPAA) template. Frist The Al foil (99.99%) was electropolished using a mixture of H_3PO_4 and H_2SO_4. The electropolished Al foil was anodized in 0.3 M oxalic acid at 40 V for 2 h. After removing the alumina layer, the 2nd anodization was carried for 5 min under the conditions and followed by the barrier-thinning process by a successive drop of the DC voltage at a rate of 0.1 V/s [13-15]. The pores of the NPAA were widened by immersing the NPAA into 6 wt% H_3PO_4 for 70 min. Au sputtering was carried out on the NPAA templates by r.f. magnetron sputtering at 65W for 1 and 2 min. The target-substrate distance was fixed at 100 mm. The Ar gas pressure was fixed at 5 mTorr during the deposition at room temperature.

The morphological parameters are characterized by field emission scanning electron microscopy (FE-SEM, JSM-7500F/JEOL). The optical properties of the uncoated and coated membranes were characterized using UV-Vis-NIR Perkin Elmer (Lambda 990) spectrophotometer.

3. Results and discussion

3.1 General NPAA morphological Properties

The topographical and morphological properties of the blank NPAA template were characterized through the FESEM images. Figure 1 shows FESEM images of NPAA template anodized for 5 min and pore widened for 70 min; (a) a cross-sectional FESEM and (b) top view FESEM. Figure 1(a) clearly shows the fabrication of highly ordered and homogenous circular nanopores in regular hexagonal nanoarrays. The diameter of the circular pores is 80 ± 2 nm. The average value of the distance between the pores, interpore distance, is ~ 100 nm. The density of the pores was estimated to be ~ 1.2×10^{10} cm^{-2}. The cross-sectional FESEM, Fig.1(b), shows smooth pores with a circular cross-section. The height of pores from the Al substrate is 410.9±8 nm. This means that the rate of growth of the alumina pores was 82.18±1.6 nm/min. The EDX spectrum of the NPAA membrane is shown in the inset of Figure 1(a). Only Al and O signals are observed, which confirm the high purity of the prepared NPAA membrane.

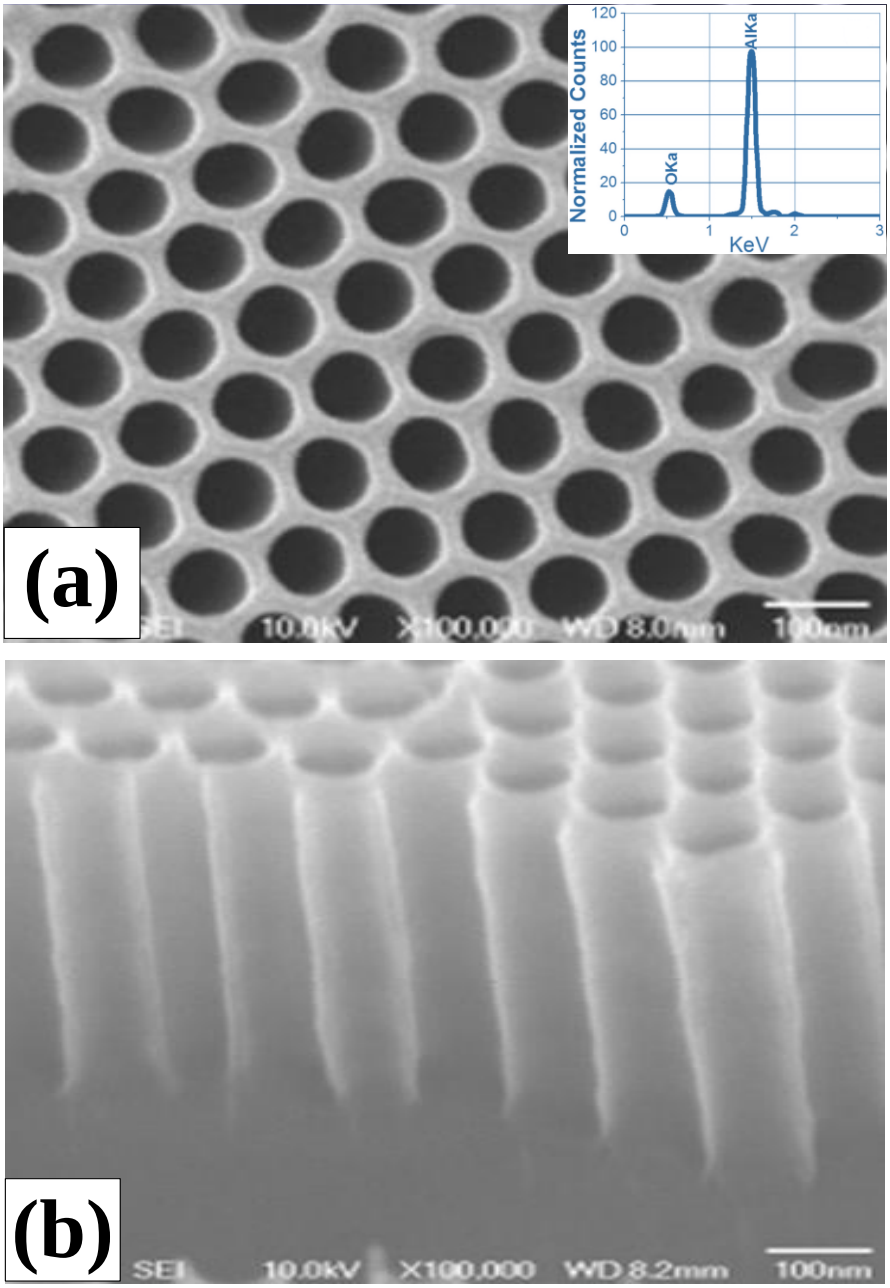


Figure 1. FESEM images of NPAA template fabricated using a two-step anodization process with 5 min second anodization followed by a 70 min pore widening process (a) top view and (b) cross-sectional view.

3.2 Morphological properties of Au/NPAA template

Figure 2 shows the FESEM image of the NPAA membrane sputtered with Au for 2 min. This image shows the growth of hexagonal nanoarray of six Au nanoparticles around each pore of the NPAA template. The narrow gap between each two successive Au particles is varied up to ~ 24 nm. Of course, the control of the Au particle size and the distance between the successive particles is sputtering time-dependent. As the sputtering time decreased, the Au nanoparticle size decreased and the gaps between the particles increased. The average Au deposition rate is

~ 14.7 nm/min. Also, the NPAA pore diameter is decreased due to the growth of Au nanoparticles on the surfaces of pore walls as seen from the inset image of Fig.2. The size distribution of these Au nanoparticles is very narrow and the density of these Au nanoparticles is very high. Generally, Au nanoparticles are grown perpendicularly on the NPAA substrate with a pronounced homogeneity in their size, shape, and density. The improved morphological properties of the Au nanoarrays could be ascribed to the homogeneity and quality of the used NPAA as a template for the deposition of plasmonic nanoarrays as discussed previously in the literature [16]. So, it is highly expected that the designed Au/NPAA can be used for different applications including catalytic systems, imaging- and diagnostic-based applications, and sensors. Hence we motivated to study the optical properties of NPAA and Au/NPAA templates.

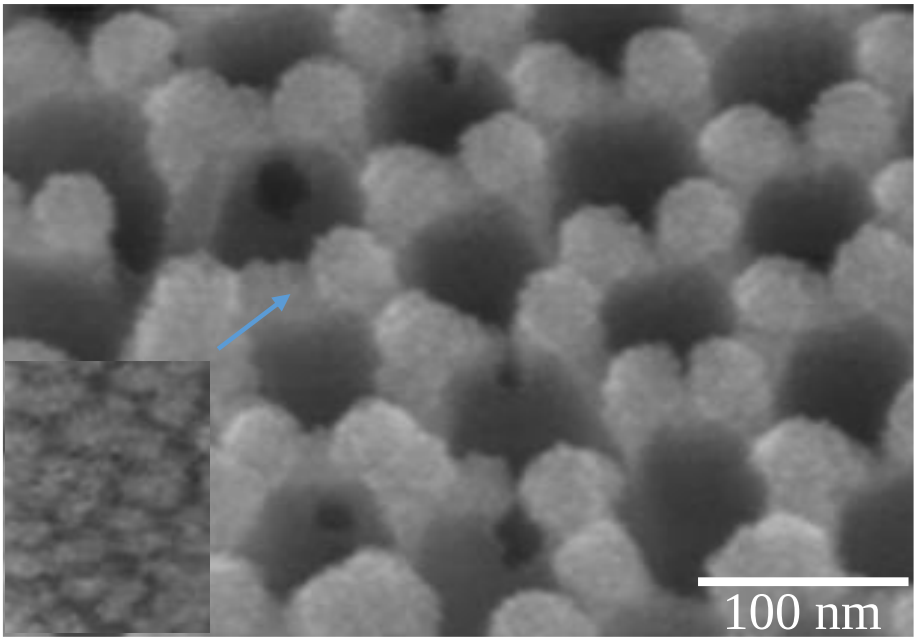


Figure 2. FESEM image NPAA template sputtered with Au for 2 min. The inset shows an enlarged FESEM image for Au nanoparticles on the pore wall.

3.3 Optical properties of NPAA and Au/NPAA templates

Figure 3 shows the reflectance spectra of uncoated and Au-coated NPAA templates for 1 min and 2 min. The reflectance spectra were measured within the spectral range 400-1000 nm using UV-Vis-NIR Perkin Elmer (Lambda 990) spectrophotometer. The main observed feature in the measured spectra indicates the interference between the reflected rays from Al/NPAA, NPAA/Au, and Au/air interfaces. So, well-set crests and troughs are observed. However, many other interesting features can be observed over the investigated spectral range. First, The reflectivity of the Au/NPAA templates is lower than the reflectivity of the uncoated NPAA template and change smoothly over the studied spectral range. Second, the reflectivity of the 1

min Au coated sample is blue-shifted and as the sputtering time increased the reflectivity increased and the reflection spectrum is redshifted. Thirdly, the saturation and the strength of the interference fringes of the Au-coated NPAA templates are highly improved relative to that of the blank NPAA.

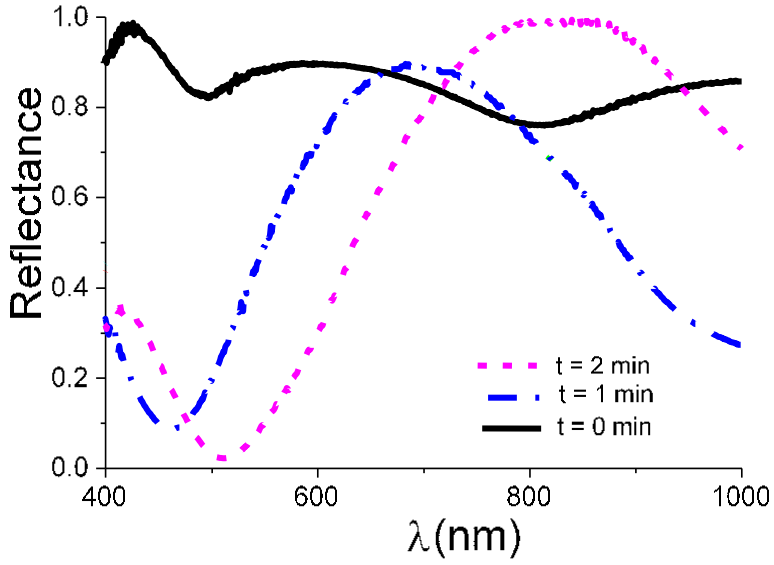


Figure 3. Reflectance spectra of uncoated, 1 min Au-coated, and 2 min Au-coated NPAA templates.

A potential explanation behind this behavior is proposed as follows. Figure 4(A) shows a schematic chart of the surface plasmon resonances coupling that produced between closest nanoparticles. This coupling causes much stronger surface plasmon resonance (SPR) modes as mentioned in literature [17, 18]. Figure 4(B) shows a schematic diagram of the deposited Au nanoarrays around each pore and on the inside pore walls. This figure also shows the generated surface SPR modes from Au/NPAA interfaces. The created surface plasmon at the Au/NPAA interface can go about as another exciting source because it can propagate into the Al_2O_3 and Al layers [19]. Here, the Au nanoarrays possess narrow gaps < 24 nm, so strong coupling, and intense SPR can be generated. For whatever length of time that the Au nanoparticles of the nanoarray come up short on a direct conductive interlink, their closeness creates negative dipole-dipole coupling energy that leads to the SPR redshift [20]. This redshift depends upon the Au particle size and the gap between the Au nanoparticles. At 1 min Au coating, the SPR wavelength is centered at 466 nm and shifted to 512 nm by expanding the deposition time to 2 min (jump 46 nm). Also, the wavelength at the maximum reflectance (λmax) is increased from about 700 to 850 nm by increasing the Au deposition time from 1 to 2 min. Thus, enormous optical fields are amassed in the leading connections between the particles of the nanoarray. Our data agree well with the previously reported data in the literature [20-22] that anomalous

enhancement impacts due to SPR and coupling can happen between the contacting surfaces of Au nanoarrays. As shown in Fig.4 (C, D), the Au coated NPAA templates showed brilliant colors with amazing saturation due to the SPR enhancement of optical interference. The photographs showed brilliant structural colors due to the strong light mirroring at certain wavelengths. As shown the colors of the samples are mostly depending on the coating time and changed from red to brown as the time increased from 1 to 2 min.

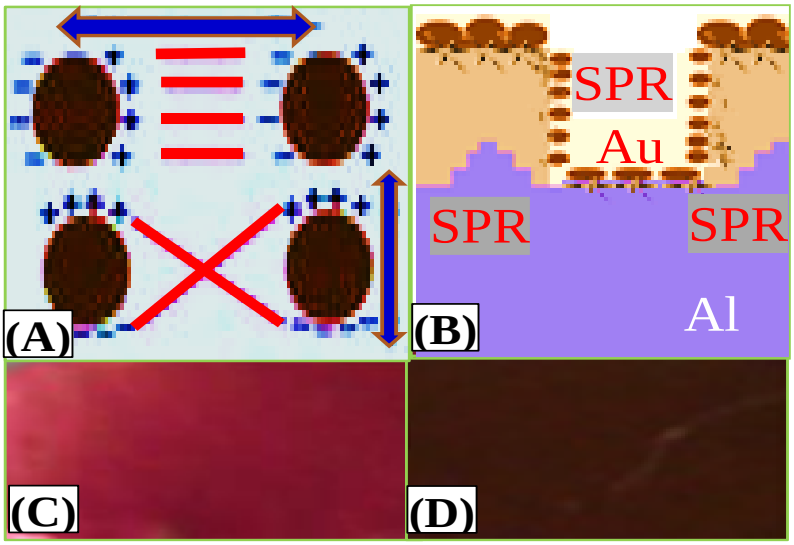


Figure 4. (A) Schematic diagram of SPR coupling modes, (B) SPR modes of Au/NPAA template, and Photographic images of (C) 1 min Au/NPAA and (D) 2 min Au/NPAA.

4. Conclusion

NPAA templates with 100 nm inter-pore distance have been fabricated by modified two-step anodization. Hexagonal Au nanoarrays have been deposited on the top surface of the NPAA template using r.f. magnetron sputtering. Simultaneously, the interior surfaces (walls) of the pores have been decorated with fine Au nanoparticles. The Au-coated templates showed brilliant and statured structural colors due to the SPR enhancement of NPAA optical interference. The SPR wavelength is increased from 466 nm to 512 nm by increasing the Au deposition time from 1 to 2 min. Because the used technique for the design of these nanostructures is simple, fast, inexpensive, and reproducible, therefore, the designed Au/NPAA nanoarrays can serve as key materials for nanophotonic and nanoelectronics devices.

5. References

[1] J. W. Diggle, T. C. Downie, and C.W. Goulding. Anodic oxide films on aluminum. Chem. Rev. 69 (1969) 365-405.

[2] Y. Yamamoto, N. Baha, and S. Tajima. Coloured materials and photoluminescence centres in anodic film on aluminium. Nature 289 (1981) 572-574.
[3] Y. Du, W.L. Cai, C.M. Mo, J. Chen, L.D. Zhang and X.G. Zhu. Preparation and photoluminescence of alumina membranes with ordered pore arrays. Appl. Phys. Lett. 74 (1999) 2951
[4] G. Wang, C. –W. Wang, Y. Li, W. –M. Liu, Thin Solid Films 516 (2008) 6789 -6798.
[5] X. H. Wang, T. Akahane, H. Orikasa, T. Kyotani, and Y. Y. Fu. Brilliant and tunable color of carbon-coated thin anodic aluminum oxide films. Appl. Phys. Lett. 91(2007) 011908.
[6] C. Hsu, and H. H. Liu, Optical behaviours of two dimensional Au nanoparticle arrays within porous anodic alumina. Journal of Physics: Conference Series. 61 (2007) 440-444
[7] S. L. Pan, D. D. Zeng, H. L. Zhang, H. L. Li. Preparation of ordered array of nanoscopic gold rods by template method and its optical properties. Appl. Phys. A 70 (2000) 637-640.
[8] K. L. Kelly, E. Coronado, L. L. Zhao, and G.C. Schatz. The Optical Properties of Metal Nanoparticles: The Influence of Size, Shape, and Dielectric Environment. J. Phys. Chem. B 107 (2003) 668-677.
[9] G. A. Wurtz, W. Dickson, D. O'Connor, R. Atkinson, W. Hendren, P. Evans, R. Pollard, and A.V. Zayats. Guided plasmonic modes in nanorod assemblies: strong electromagnetic coupling regime. Optics Express 16 (2008) 7460-7470.
[10] X. Zhang, B. Sun, H. Guo, N. Tetreault, G. Giessen, R. H. Friend. Large-area two-dimensional photonic crystals of metallic nanocylinders based on colloidal gold nanoparticles. Appl. Phys. Lett. 90 (2007) 133114.
[11] S. Nie, S. R. Emory. Probing Single Molecules and Single Nanoparticles by Surface-Enhanced Raman Scattering. Science 275(1997) 1102-1106
[12] J. –Y. Chu, T. –J. Wang, Y. –C. Chang, M. –W. Lin, J. –T. Yeh, and J. –K.Wang. Multi-wavelength heterodyne-detected scattering-type scanning near-field optical microscopy. Ultramicroscopy 108 (2008) 314-319.
[13] M. Shaban, H. Hamdy, F. Shahin, Joonmo Park, and S.-W. Ryu. Uniform and Reproducible Barrier Layer Removal of Porous Anodic Alumina Membrane. J. Nanosci. Nanotechnol. 10 (2010) 3380-3384.
[14] M. Shaban, M. Ali, K. Abdelhady, A.A. P. Khan, H. Hamdy, Hexagonal arrays of Pt nanocylinders on the top surface of PAA membranes using low vacuum sputter coating technique, Vacuum 161 (2019) 259-267.
[15] M. Shaban, Morphological and Optical Characterization of High Density Au/PAA Nanoarrays, Journal of Spectroscopy 2016 (2016) 8 page [Article ID 5083482].
[16] C. E. Rayford, G. Schatz, and K. Shuford. Optical properties of gold nanospheres. Nanoscape 2 (2005) 27-33.
[17] M. Shaban, H. Hamdy, F. Shahin, and S.-W. Ryu. Strong surface plasmon resonance of ordered gold nanorod array fabricated in porous anodic alumina template. J. Nanosci. Nanotechnol. 10 (2010) 3034-3037.
[18] P.R. Evans, R. Kullock, W. R Hendren, R. Atkinson, R. J.Pollard, and L. M. Eng. Optical Transmission Properties and Electric Field Distribution of Interacting 2D Silver Nanorod Arrays. Adv. Funct. Mater. 18 (2008) 1075 -1079.
[19] N. Felidj, J. Aubard, G. J. Levi, R. Krenn, G. Schider, A. Leitner, and F. R. Aussenegg. Enhanced substrate-induced coupling in two-dimensional gold nanoparticle arrays. Phy. Rev. B 66 (2002) 245407.
[20] T. Atay, J-H. Song, and A. V. Nurmikko Strongly Interacting Plasmon Nanoparticle Pairs: From Dipole−Dipole Interaction to Conductively Coupled Regime. Nano Letters 4 (2004) 1627-1631.
[21] M. Shaban, Ashour M. Ahmed, Ehab M. Abdel Rhman, and Hany Hamdy, "Tunability and Sensing Properties of Plasmonic/1D Photonic Crystal, Scientific Reports 7 (2017) 41983.
[22] M. Rabia., M. Shaban, A. Adel and A.A. Abdel-Khaliek, Effect of plasmonic au nanoparticles on the photoactivity of polyaniline/indium tin oxide electrodes for water splitting. Environmental Progress and Sustainable Energy 38 (2019) 1-8.

Effect of Pico Second Laser on Structural and Photoluminescence Properties of Aluminum Doped Zno Film Prepared by Sol Gel Method

M. Naziruddin Khan,
Department of Physics, Faculty of Science, Islamic University of Madinah, KSA
mnkhan_phy@yahoo.com, mnkhan@iu.edu.sa
and
Abdullah Almohammedi
Department of Physics, Faculty of Science, Islamic University of Madinah, KSA
M.A Majeed Khan,
King Abdullah Institute for Nanotechnology, King Saud University, Riyadh, KSA

Abstract: Aluminum (Al) doped zinc oxide (Al-doped ZnO) thin films with different Al concentrations were prepared using the sol-gel method and the spin coating technique. The films were irradiated by high power pico-second pulsed laser source of 355 nm with different energy. The laser irradiated Al-doped ZnO films were then characterized by X-ray diffraction, scanning microscopy, UV–Vis absorption and photoluminescence techniques. The diffraction pattern indicates increase in crystallinity of ZnO thin films due to the influence of Al doping and laser irradiation. The change in surface morphology of ZnO and Al-doped ZnO films are confirmed by SEM. The absorption spectra of the films show that slightly change in band gap of pure and Al-doped ZnO after the laser treatment. No significant change in band gap with different Al doping percent under laser treatment. Significant influence in photoluminescence (PL) and spontaneous properties of the pure ZnO and Al-doped ZnO thin films by laser energy was observed. The PL and spontaneous intensity of Al-doped ZnO films decreases upon increasing the laser energy which is attributed to effect of nanostructured of the film.

Keywords: ZnO films, Aluminum, SEM, Absorption, Emission, Laser source

تأثير ليزر بيكو الثاني على الخصائص الإنشائية والتلألؤ الضوئي لفيلم أكسيد الزنك المخدر بالألمنيوم المحضر بطريقة المحلول الهلامي

الملخص: تم تحضير أغشية رقيقة من الألومنيوم (Al) ومخدر أكسيد الزنك (Al-doped ZnO) بتركيزات مختلفة من الألومنيوم باستخدام طريقة المحلول الهلامي وتقنية الطلاء الدوراني. تم تشعيع الأغشية بمصدر ليزر نبضي عالي الطاقة يبلغ ٣٥٥ نانومتر مع طاقة مختلفة. تم بعد ذلك تمييز أفلام مخدر من أكسيد الزنك المشعة بالليزر بتقنيات حيود الأشعة السينية والمسح المجهري وتقنيات امتصاص UV-Vis والتلألؤ الضوئي. يشير نمط الحيود إلى زيادة في تبلور أغشية أكسيد الزنك الرقيقة نتيجة لتأثير المنشطات Al وتشعيع الليزر. تم تأكيد التغيير في التشكل السطحي لأفلام أكسيد الزنك ومخدر أكسيد الزنك بواسطة. SEM تظهر أطياف الامتصاص للأغشية تغيرًا طفيفًا في فجوة النطاق من أكسيد الزنك النقي و المخدر بعد العلاج بالليزر. لا يوجد تغيير كبير في فجوة النطاق مع نسبة المنشطات Al المختلفة تحت العلاج بالليزر. لوحظ تأثير كبير في التلألؤ الضوئي (PL) والخصائص التلقائية لأغشية أكسيد الزنك النقية ومخدر أكسيد الزنك بواسطة طاقة الليزر. تنخفض كثافة PL والشدة التلقائية لأغشية مخدر أكسيد الزنك عند زيادة طاقة الليزر التي تُعزى إلى تأثير البنية النانوية للفيلم.

1. Introduction

Zinc oxide (ZnO) based nanostructured materials have been considered as a self-structured growth into different such as nanorods, nanowire etc., which are a promising material for different optoelectronic devices, in particular light emitting diodes (LEDs) and laser diodes (LDs) [1-4]. Moreover, ZnO has a wide band gap with energy of 3.37eV and a large exciton binding energy of 60 meV at room temperature, commonly considered as an efficient UV emitter [5]. During the past years, the different properties of Zinc Oxide (ZnO) based transparent conducting oxide (TCO) have been studied and proposed to be favorable materials for applications of LEDs, optical waveguides, piezoelectric transducers, gas sensors and transparent conductive electrodes in solar cells [6-8].

On the other hand, the metals-doped ZnO were extensively interested for their enhancement of properties due to native defects existent in its structure. It can reduce the high electrical resistivity of ZnO by doping metals such as aluminum (Al), boron (B), gallium (Ga), or indium (In) [9-12]. Aluminum (Al) is considered as one of the suitable dopant among the metals due to its smaller ionic radius and lower cost. Addition of Al in zinc oxide materials can be improved in their electrical conductivity by replacing the Al ions. Several researches were endeavored to improve in electrical and optical property of Al-doped ZnO thin films and powder using different methods such as spray pyrolysis [13], sol–gel method [14], precipitation [15] and hydrothermal processes [16]. Traditional post-annealing in a furnace at high temperatures in excess of 700 ^{0}C has undesirable effects on device fabrication, which give rise to the high cost of the device fabrication [17]. In view of these, our group studied on the properties of nanocrystals Al-doped ZnO that noticed a blue shift in energy band gap and UV emission due to the carrier concentration and degree of crystalline order induced by Al-donor [18].

In addition, researches are extensively interested to engineer and improve in the different properties of the TCOs material for various applications especially in optoelectronic devices. For this, laser irradiation technique has shown as a powerful tool for surface transformation and physical properties of certain functional materials [19-20]. Because the laser irradiation technique offers some advantages over the traditional annealing method in terms of less in thermal exposure, local-heating, reduced effect on the substrate temperature, and easy manageable [21]. Further, laser annealing technique can yield the high quality Al-doped Zinc oxide (AZO) films at low temperature for the fabrication of solar cell and liquid crystal display (LCD) [22]. Some reports of laser irradiation effect on the optical band gaps [23-24] and

photoluminescence (PL) [25-28] of non-doping ZnO nanostructured films were appeared in literature. Many more works to develop the high quality ZnO for the improvement in electrical conductivity using laser irradiation technique have been reported [29-30]. Further, the electrical, surface morphology, optical transmission and gas sensing properties including resistivity and optical band-gap of Al-doped ZnO films were investigated [31-33].

Though the laser irradiation studies on the electrical properties of Al-doped ZnO thin films were reported, there is still limited works on the optical, infrared (IR), structural and photoluminescence (PL) properties of nanostructured Al-doped ZnO thin films. Recently, we have reported the effect on optical properties of pure ZnO films by laser irradiation [34]. In continuation of previous work, studied on the irradiation effect of Al-doped ZnO thin films using different energy of a high power pico second laser (355nm). The main interest of 355 nm laser is UV source that can impact on the surface of thin film. The influence of laser energy on the different properties of the Al-doped ZnO thin films were investigated.

2. Materials and Method

2.1. Preparation

The ZnO solution was prepared by dissolving the zinc acetate dehydrate [ZnAc:($Zn(CH_3COO)_2$ $2H_2O$)] in 2-proponal and diethanolamine (DEA, $C_4H_{11}NO_2$). In the beginning, ZnAc was dissolved in 2-proponal and then added DEA to increase the solubility. The corresponding molar ratio of DEA/ZnAc to solution was taken as 1 with 0.5 M concentration. In addition, dopant agent was taken as aluminium acetate basic hydrate [(CH3CO2) AlOHH2O] for Al-doped ZnO precursor solution. Al-doped ZnO thin films were prepared using different concentration of Al. The molar ratios of Al/Zn were taken as Al^{+3}/Zn^{+2} = 0.5 % and 1.0 %, by mixing the aluminium acetate basic hydrate into the precursor solution used for ZnO. The final solution was stirred at 700 ^{0}C for 2 hr to produce homogenous solution. The size of glass substrates was 10 x 10 mm^2 and cut and cleaned with ethanol for 10 min before used. Cleaning of substrate was performed in an ultrasonic cleaner and finally cleaned with deionized water and dried. The gel solution was dropped onto cleaned glass substrate at 3000 rpm for 30 s using a spin coater (Laurell EDC-650-23B). After the deposition by spin coating, the films were preheated at 2500 ^{0}C for 10 min on hot plate to evaporate the solvent and remove organic residuals. Finally, the films were annealed 4000 ^{0}C for 2 hr.

2.2. Characterizations

All prepared thin films were irradiated using 355 nm laser source from a mode-locked and Q-switched Nd: YAG Pico Second Tunable Laser System (LS-2151, LOTTIS II, Belarus). The

355 nm wavelength was generated with third harmonic generation from fundamental wavelength 1064 nm. Pulse duration of the fundamental beam was 70-80 ps. The pulsed repetition rate of the laser was 15 Hz. The maximum output pulse energy at 355 nm with 15Hz is 21 mJ. In the experiment, the very short laser pulse of energy fluence at 10 mJ/cm2, 15 mJ/cm2 and 20 mJ/cm2. The laser exposure time for all samples was about 3 min at 6 mJ, 9 mJ and 12 mJ energy of 355 nm laser source was used under the below threshold ablation and varied the peak power of laser.

The sample was positioned without focus the beam and keeping all other parameters constant except the laser energy. The exposure area of the sample with the laser beam are fixed during the laser irradiation. The laser energy is increased to see the laser influence on surface of the films. The diameter of laser beam was 6 mm and exposed on the sample area since the position of all samples was fixed.

The optical absorption, emission and Infrared spectra (IR) of the films before and after laser irradiation were measured in UV-Visible-NIR spectrophotometer (670, JASCO), Lumina Fluorescence spectrophotometer (Thermo Scientific,) and Fourier Transform Infrared Spectroscopy (FTIR) (Perkin Elmer) respectively. XRD images of the ZnO and Al-doped ZnO films was investigated using X-ray diffraction (XRD, PANalytical X'Pert) with Cu Kα radiation (λ=0.154 nm) operated at 45 kV and 40 mA. The surface morphology of Al-doped ZnO films before and after laser irradiation was studied by a Field Emission Electron Microscope (FESEM).

3. Results and Discussion

3.1. Structural analysis: The influences of laser energy on the structural property of pure ZnO and Al-doped ZnO films were investigated. The characterization of crystals and molecular structures of the ZnO and Al-doped ZnO films were conducted by X-ray diffraction and FTIR techniques. The XRD feature of pure ZnO and Al-doped thin films before and after the laser treatment at different energy is shown in Figure 1. The observed diffraction peaks including the prominent (101), (002) and (101) of ZnO correspond to a hexagonal wurtzite crystal structure. These observed peaks are in accordance with the pristine and the irradiated films which is confirmed from (JCPDS card No. 36-1451). There are no additional peaks observed which correspond to other crystalline phases or aluminum. The observed peaks are generally broadening which indicate the smaller sizes of the particle. There is slightly deteriorated in intensities of the pure ZnO film after laser irradiation with time. As clearly seen in Figure 1 that pattern of ZnO is slightly broaden when Al is contained within ZnO. The reason of

broadening the peaks with Al containing ZnO may attribute to the effect on size of the particles. Intensities of Al-doped ZnO after laser exposure is influenced over the irradiation time but the rate of decrease in their intensities are different from the pure ZnO film. It means that induce of laser energy power on the structure of Al-doped films are less influenced as compared to pure ZnO due to Al involvement. The diffraction pattern indicates that increase in crystallinity of ZnO thin films due to the influence of Al doping and laser irradiation. The reason may possible and attribute to the addition of Al and absorption of heat radiation from laser source promote grain growth of the films [19].

The infrared spectra of the ZnO and Al-doped ZnO films was investigated by IR over 500-4000 cm^{-1} region. The IR peak positions of ZnO film in the 2000-4000 cm^{-1} region are slightly shifted as compared to the relative peaks of Al-doped ZnO films. When the laser irradiated on ZnO and Al-doped ZnO films, the relative intensities of IR features are decreased with increasing the laser energy as shown in Figure 2 (i, ii, iii). Doping of Al in ZnO induces to the relative intensity. When the Al doping percentage increases, the relative intensity of ZnO is slightly decreased. For instance, the absorption intensity of features around at 3570 cm-1, 2916 cm-1, 2852 cm-1 and 2107 cm-1 (shoulder) are slightly reduced due to laser irradiation with increasing the time. For instance, the two bands of ZnO at round 2852 cm-1 and 2916 cm-1 are noticeably decrease in their relative intensities with respect to irradiation time.

These two peaks of very weak intensities at 2852 and 2916, cm^{-1} are due to C–H stretching vibrations of alkane groups [35]. There is small influenced of Al to the peaks of ZnO. The observed peaks correlate well with the observed frequencies of the $C–H_2$ symmetric stretch (2855 cm^{-1}), $C–H_2$ asymmetric stretch (2926 cm^{-1}), and $C–H_3$ asymmetric stretch (2962 + 10 cm^{-1}) of saturated hydrocarbons, respectively as reported in [36].

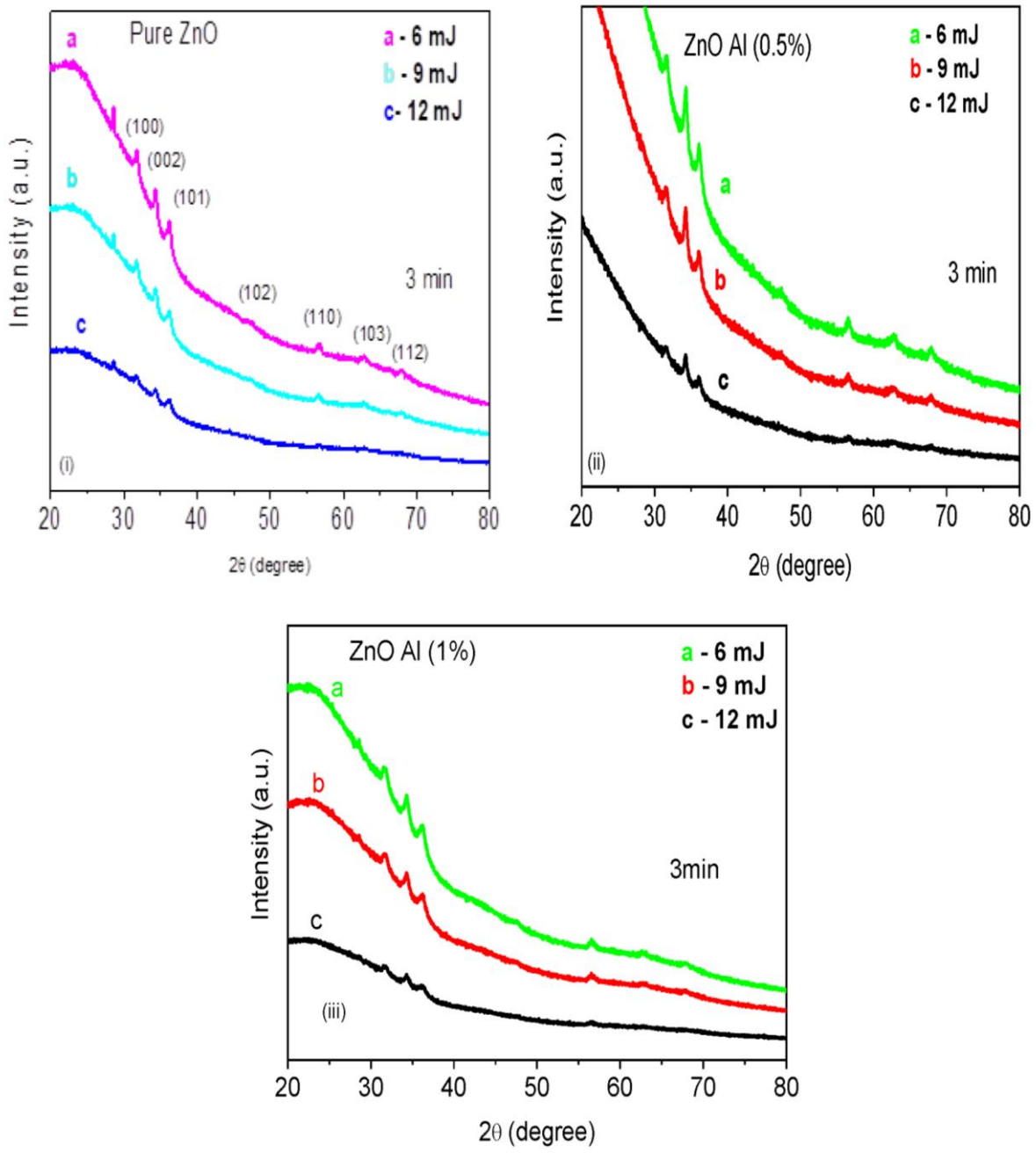


Figure 1: XRD spectra of (i) pure ZnO and (ii and iii) Al-doped ZnO after irradiation at different laser energy

Overall the change in relative intensities of Al-doped ZnO films after laser irradiation may attribute to the influence on structure of Al-doped ZnO due to laser energy.

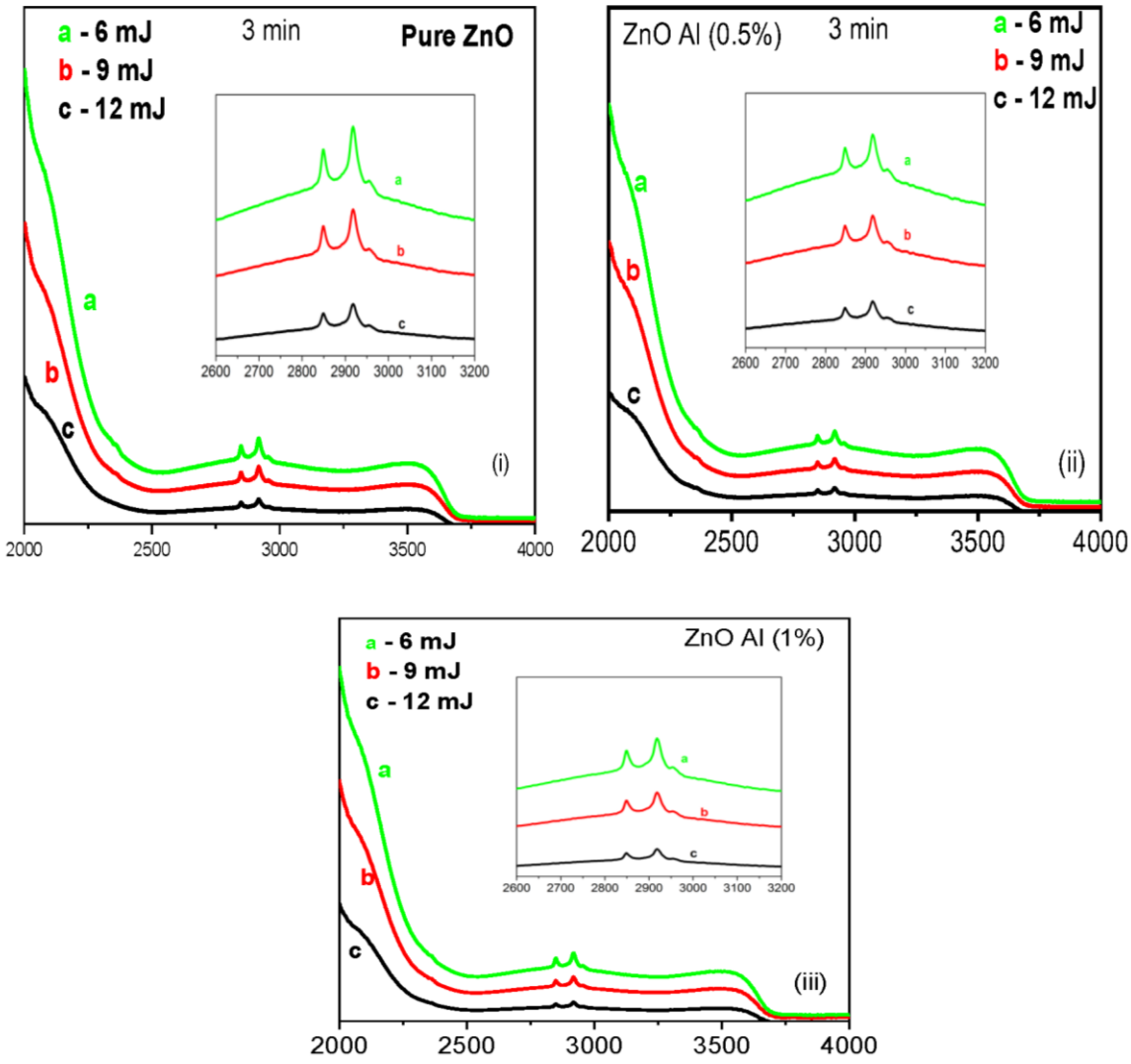


Figure 2: IR spectra of (i) Pure ZnO and (ii and iii) Al-doped ZnO after irradiation at different laser energy

The existence of ZnO nanorods in the surface of film is seen in the image of surface morphology of ZnO film is displayed in Figure 3 (i). The SEM images of Al-doped ZnO films before and after laser irradiation is shown in Figure 3 (ii-iii). The presence of ZnO nano rods with Al materials in film is confirmed by the EDX spectrum as shown in Figure 3 (iv). The surface of Al-doped ZnO film before laser treatment shows ZnO nano rods exist inside Al surface layer. After the laser treatment on the surface, dispersal of ZnO nanorods in the Al layer is detected with a condense surface. It means that after the laser exposure on the surface, the distribution of ZnO nano rods in the Al containing surface is remarkably changed with their location i.e., majority of nano rods becomes agglomerate and overlap. It seems to be loosely binding with Al. It suggests that when the laser beam fall on surface, energy could penetrate and deform the uniformity of Al-doped ZnO surface that lead to aggregate the nano rods as

shown in Figure 3 (iii). The agglomeration of nano-rods may effect that when laser light interacts with the surface of Al-doped ZnO, laser energy could induce to defect to Al and ZnO nano rods. Therefore, it may lead to cause the length of ZnO rod become shorten.

The morphological particle size of Al-doped ZnO was clearly seen in figure 3 that the average sizes was significant decreased after laser irradiation. Similar result was reported the effect on sizes of Al-doped ZnO as the laser energy exposed [37].

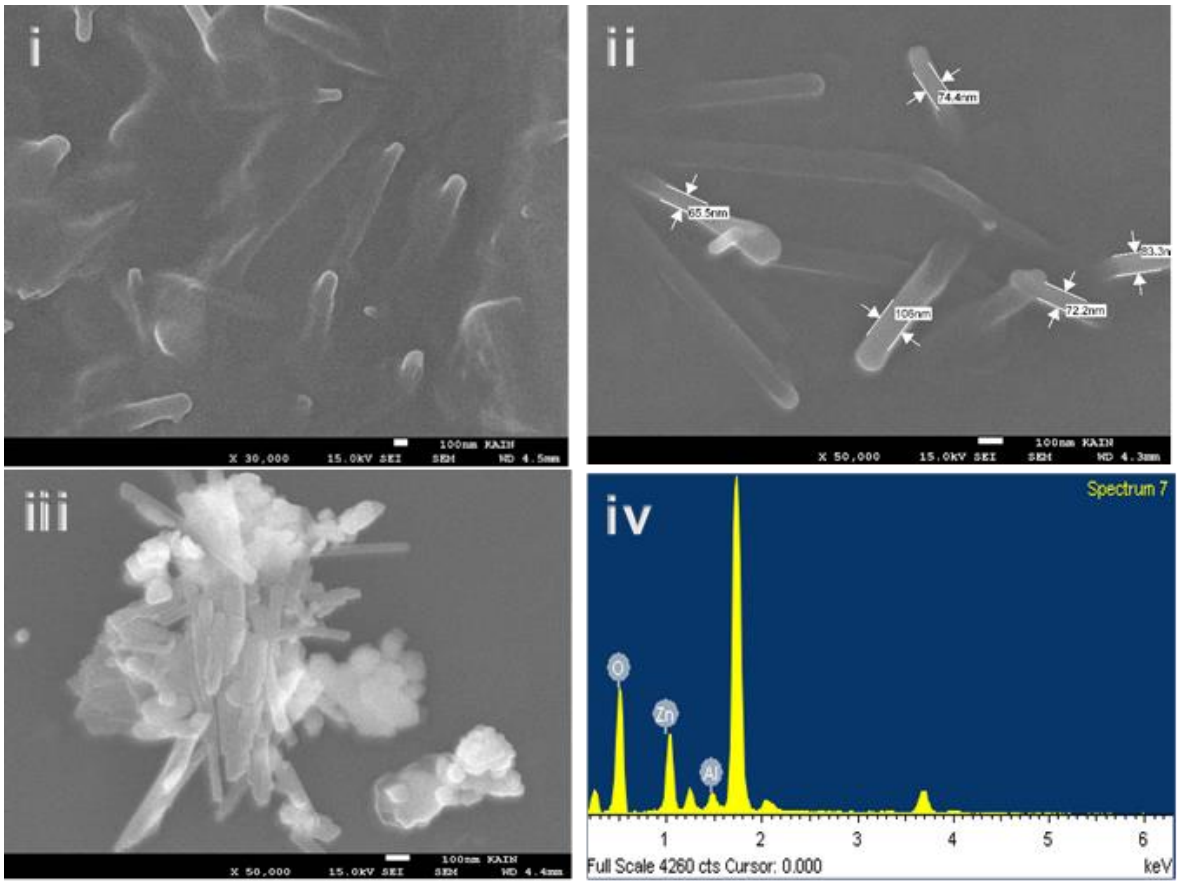


Figure 3. FESEM images of (i) ZnO [34],
(ii-iii) Al-doped ZnO film before and after laser irradiation, (iv) EDX spectrum

3.2. Absorption property

The absorption property of pure ZnO and Al-doped ZnO films was studied after the laser treatment at different laser energy. The laser irradiation influence in the absorbance of the samples were recorded. The band gap energies of the films were also determined using the absorption spectra. The typical absorption spectra for pure ZnO and Al-doped ZnO films after the laser irradiation are shown in Figure 4 (i-iii). The absorption edge for pure ZnO and Al-doped ZnO films shows slightly sharp after irradiation the films. The peaks are blue shifted from the corresponding bulk value (340 nm) as listed in table 1.

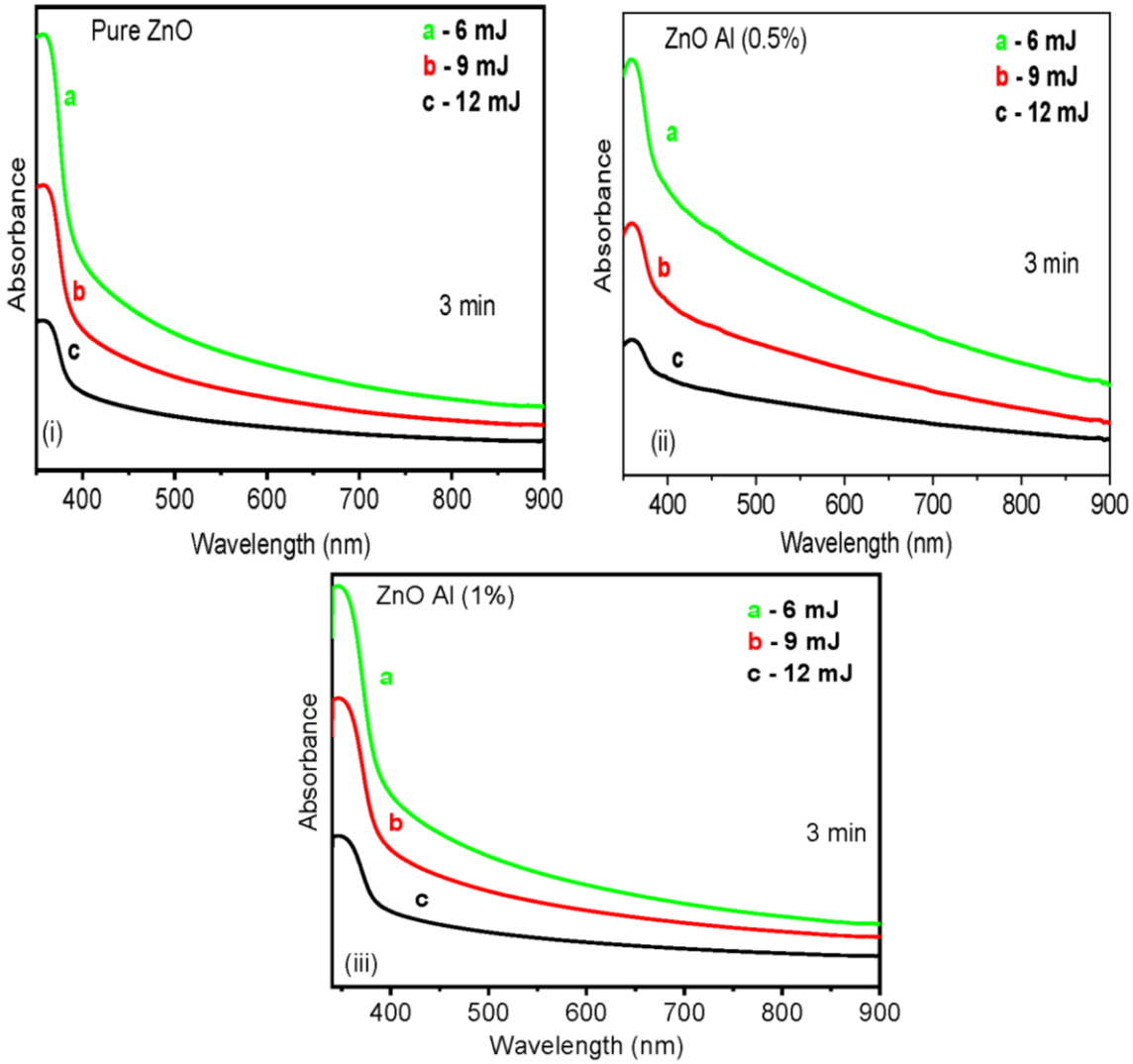


Figure 4: Absorption spectra of (i) pure ZnO and
(ii and iii) Al-doped ZnO after irradiation at different laser energy

The absorbance of the ZnO film was affected as the pump energy increase from 6 mJ to 9 mJ and 12 mJ as shown in Figure 4 (i). There is similar pattern shown in the absorbance of Al-doped ZnO at different laser energy. The rate of decay in peaks of Al-doped film with respect to different laser energy is more inclined. As clearly seen in Figure 4 (ii-iii), absorption spectra of Al-doped ZnO with Al (0.5 %) and (1%) is observed different from the absorption spectra of pure ZnO under the laser treatment. In general, absorption peaks of ZnO is influenced when the Al included. The reason of influence is not discussed in detail since the interest of present work was focused on laser energy effects on the ZnO and Al doped materials. But, the reason of Aluminum influence in ZnO were attributed to the increase in band-gap energy, decrease in particle sizes and surface roughness when Al included as detailed in [18, 38-39].

Table 1: Absorption peak positions of Pure ZnO and Al-doped ZnO

Pump energy	Peak positions of Pure ZnO	Peak positions of Al-doped ZnO Al (0.5%)	Peak positions Al-doped ZnO Al (1%)
6 mJ	359 nm	360 nm	352 nm
9 mJ	361 nm	362 nm	352 nm
12 mJ	361 nm	362 nm	352 nm

The absorption coefficient α associated with the optical band gap for a direct band gap can be determined by Tauc plots:

$$\alpha h\nu = A\left(h\nu - E_g\right)^{1/2}$$

Where $h\nu$ is photon energy, A is the transition probability of the parameter and the optical band gap represents as E_g. The Eg value is determined from an intercept on energy axis of $h\nu$ vs $(\alpha h\nu)^2$ plot as shown in Figure 5 (i-iii) and given in table 2. It is seen from table 2, there is slightly effect in band gap of ZnO and Al-doped ZnO films under laser treatment. The energy gap values increase with the increasing of the energy fluences. No doubt, effects in band gap are observed when the Al doped into ZnO nanostructures as listed in table 2, the reason of effect may attribute to the introduction of defect levels within the band gap.

As clearly seen in table 2, there is slightly changed in energy band-gap with the variation of irradiation laser energy which may attribute to little effects in crystallite size with the thickness of the films as a result of confinement effect [35] and heat on samples etc.

Table 2: The energy band-gap (Eg) values of pure ZnO and Al-doped ZnO

Pump energy	Pure ZnO (E_g)	Al-doped ZnO Al (0.5%) (E_g)	Al-doped ZnO Al (1%) (E_g)
6 mJ	3.04 ev	2.97 ev	2.9 ev
9 mJ	3.07 ev	3.05 ev	2.99 ev
12 mJ	3.09 ev	3.08 ev	3.1 ev

There is no much significant effect on band gap of ZnO and Al-doped ZnO by laser energy but the stability of band gap is slightly higher with ZnO than the Al-doped ZnO as seen in listed table 2.

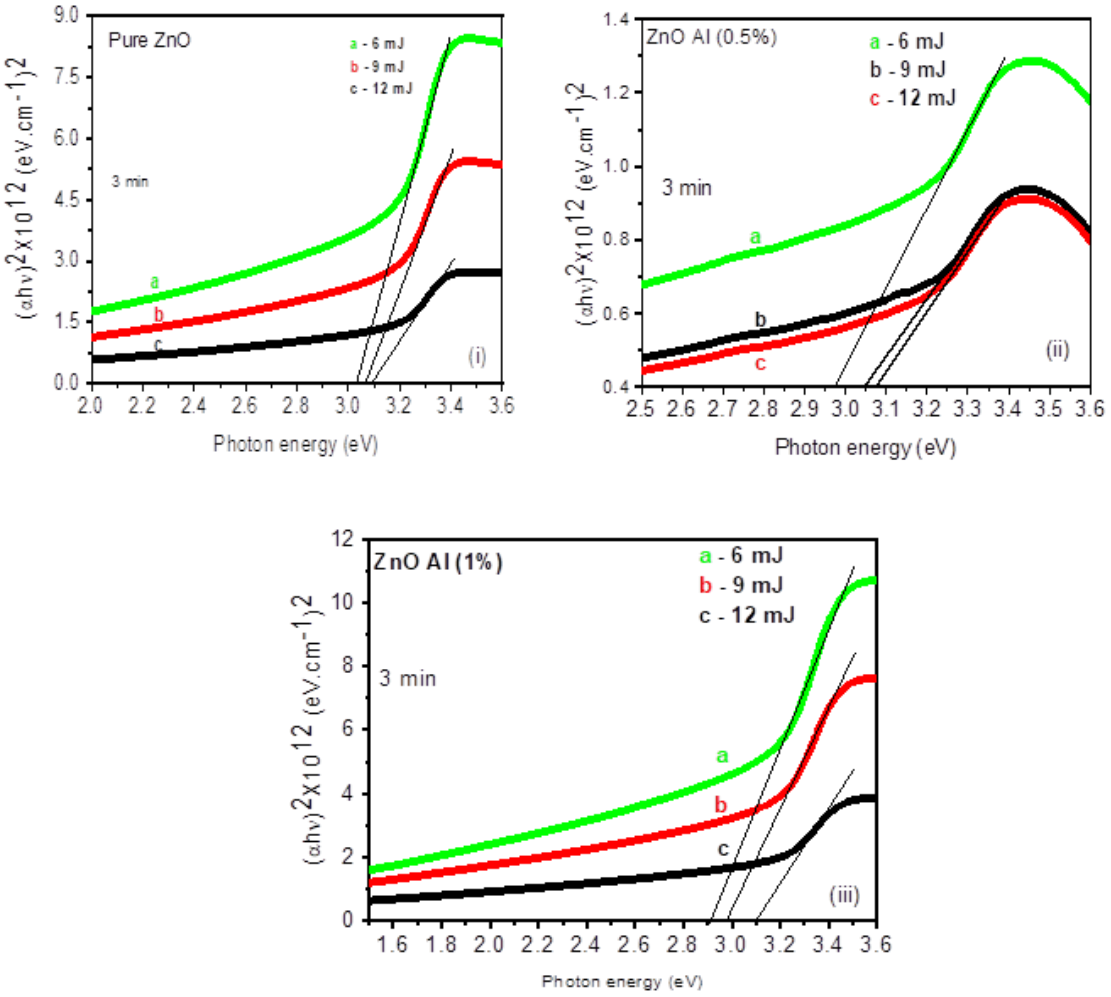


Figure 5: Energy bandgap (Eg) of (i) pure ZnO and
(ii and iii) Al-doped ZnO after irradiation at different laser energy

3.3. Photoluminescence spectra

The room-temperature photoluminescence spectra of pure and Al-doped ZnO thin films irradiated at different energy of Pico second laser source is shown in Figure 6 (i-iii). The emission spectra observed from the samples were excited at 325 nm as excitation wavelength under same environment. The ZnO sample exhibits similar UV emission band centred at 385 nm, which is attributed to band-to-band transitions, excitonic emissions [40-41], and donor-acceptor pair transitions.

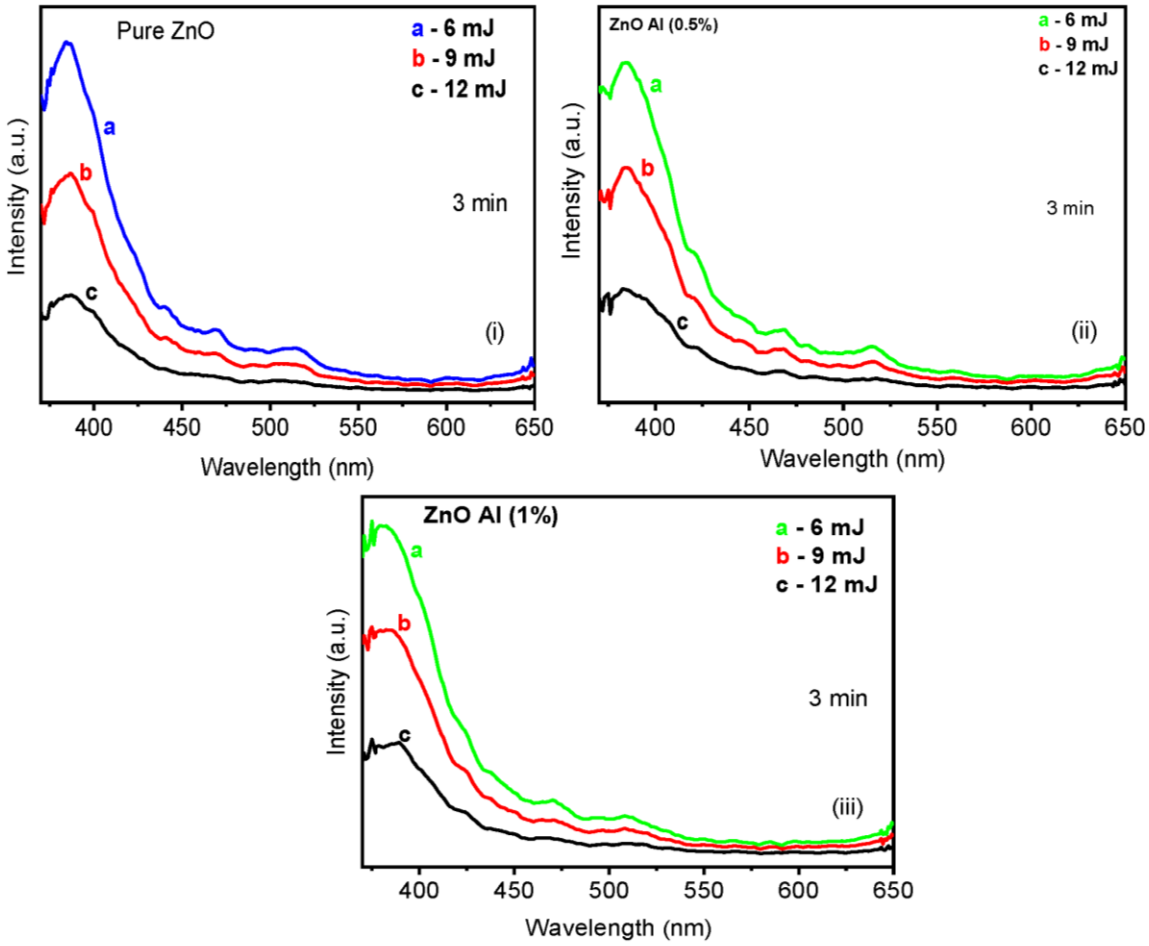


Figure 6. Photoluminescence spectra of (i) pure ZnO
and (ii and iii) Al-doped ZnO after irradiation at different laser energy.

The observed emission peak corresponds closely to calculated bandgap wavelength. The band edges remain sharp of the pure ZnO and Al-doped ZnO which correspond to emission, indicating well integrated the Al dopants into the ZnO lattice. But the effect in emission intensity of ZnO and Al-doped ZnO films are significant after the laser irradiation as shown in

Figure 6 (i-iii). The rate of decay on emission peaks of ZnO and Al-doped ZnO films is followed by the different laser energies. For instance, emission peak with band-shape of ZnO film is slightly stable as compare to Al-doped ZnO after laser exposure. Even the higher percent (Al-1%) of Al-doped ZnO film is more affected in intensity due to laser energy but remain stable in peak positions as shown in Figure 6 (ii-iii). It means that no significant effect on emission peaks of ZnO and Al-doped ZnO film due to laser energies except the intensity and band shape as seen in listed table 3. It is worth to mention that different laser energies make influence on the surface of both type of samples but there is more influenced on Al-doped ZnO. It may contribute to the integration of Al in ZnO films that laser energy more penetrate in layer of film.

In order to confirm, the emission property of both films i. e, pure ZnO and Al-doped ZnO films, spontaneous emission spectra were recorded during the laser exposure at same environment. The spontaneous emission spectra of ZnO and Al-doped ZnO films are displayed as in Figure 7 (i-iii).

Table 3: Peak positions of emission and spontaneous spectra of Pure ZnO and Al-doped ZnO after irradiation at different laser energy

	Emission peaks			Spontaneous peaks		
Samples	6 mJ	9 mJ	12 mJ	6 mJ	9 mJ	12 mJ
Pure ZnO	385 nm	387 nm	387 nm	389 nm	387 nm	387 nm
Al-doped ZnO Al (0.5 %)	384 nm	383 nm	383 nm	390 nm	389 nm	389 nm
Al-doped ZnO Al (1 %)	381 nm	383 nm	385 nm	388 nm	390 nm	390 nm

As clearly seen in the figure and table 3, the spontaneous peaks, intensities and band shape are differently observed. Though emission peaks of both type samples do not shift significant under laser irradiation but there is substantial different at peak position of spontaneous peak position from emission peak. For instance, the emission peaks of Al-doped ZnO (0.5 %) and (1 %) do not much shift upon the different pumped energy but the corresponding spontaneous peaks of the samples are significantly influenced by laser energy. The minor change in emission peaks of the pure ZnO and Al-doped ZnO films under different energy is attributed to site effects as well influence of laser energy.

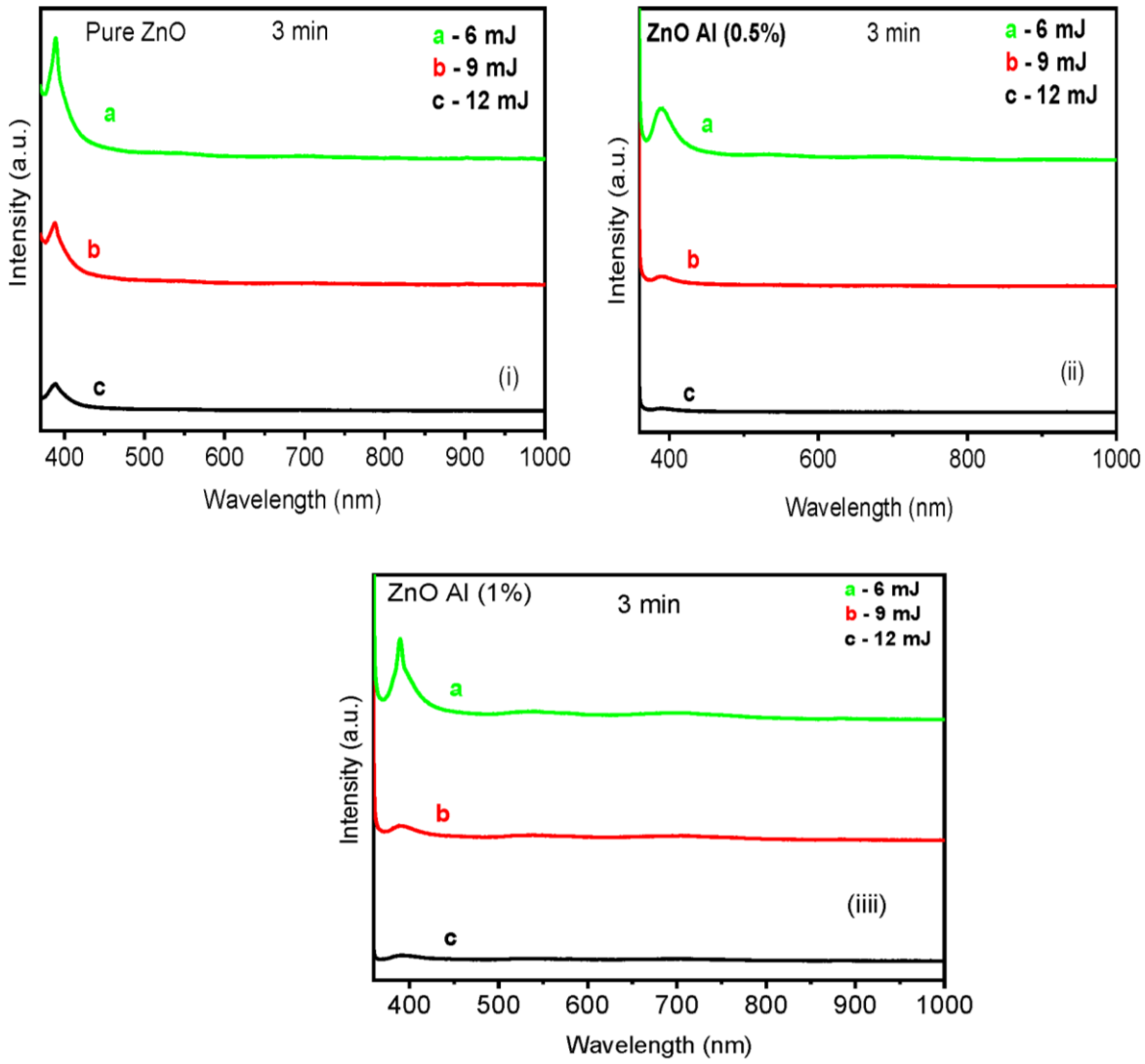


Figure 7: Spontaneous spectra of (i) pure ZnO
and (ii and iiii) Al-doped ZnO at different laser energy.

In both observation, intensities are deteriorated with respect to laser energies. The rate of deterioration on intensifies for the samples is slightly higher at Al doped sample and also depend on the percent of dopant. It may a reason to the change in grain distribution and roughness of surface in films after laser irradiation that can lead to deteriorate the intensities of emission peaks.

4. Conclusion

In present work, pure ZnO and Al-doped (0.5 % and 1 %) ZnO thin films were prepared by the sol-gel method using spin-coating technique. Laser irradiation effects on structure, optical, energy band gap and emission spectra induced by Pico second laser source were investigated.

Results reveals that laser irradiation influences on the optical properties and structure. There is no significant caused at fundamental peaks of pure and Al-doped ZnO films but substantial changed in their intensities. It is found that the band gap narrowing with changed in films due to laser power which has been attributed to the increase in the localized tail states. The observed photoluminescence spectra of pure ZnO and Al-doped ZnO thin films exhibits near-band edge (NBE) emission peaks. In addition, the PL and spontaneous intensities of the Al-doped films decreases after the laser irradiation due to defect annihilation.

References

[1] S. F. Yu, C. Yuen, S.P. Lau, W. I. Park, G.C. Yi, Random laser action in ZnO nanorodarrays embedded in ZnO epilayers, Appl. Phys. Lett., 84, 3241, 2004.
[2] T. Shimogaki, K. Okazaki, D. Nakamura, M. Higashihata, T. Asano, T. Okada, Effect of laser annealing on photoluminescence properties of Phosphorus implanted ZnO nanorods, Optics express, 20, 15248, 2012.
[3] R. Q. Guo, J. Nishimura, M. Matsumoto, D. Nakamura, T. Okada, Catalyst-free synthesis of vertically-aligned ZnO nanowires by nanoparticle-assisted pulsed laser deposition, Appl. Phys. A Mater. Sci. Process, 93, 843, 2008.
[4] X. W. Sun, B. Ling, J. L. Zhao, S. T. Tan, Y. Yang, Y. Q. Shen, Z. L. Dong, X. C. Li, Ultraviolet emission from a ZnO rod homojunction light-emitting diode, Appl. Phys. Lett., 95, 133124, 2009.
[5] X. Sun, H. Kwork, Optical properties of epitaxially grown zinc oxide films on sapphire by pulsed laser deposition, J. of Appl. Phys., 86, 408, 1999.
[6] R. Triboulet, J. Perrière, Epitaxial growth of ZnO films. Prog. Cryst. Growth Charact. Mater., 47, 65-138, 2003.
[7] C. Shaoqiang, Z. Jian, F. Xiao, W. Xiaohua, L. Laiqiang, S. Yanling, X. Qingsong, W. Chang, Z. Jianzhong, Z. Ziqiang, Nanocrystalline ZnO thin films on porous silicon/silicon substrates obtained by sol–gel technique, Appl. Surf. Sci. 241, 384-391, 2005.
[8] S.J. Jiao, Z.Z. Zhang, M. Y. Lu, D. Z. Shen, B. Yao, J. Z. Zhang, B. H. Li, D. X. Zhao, X. W. Fan, Z. K. Tang, 2006. ZnO p-n junction light-emitting diodes fabricated on sapphire substrates, Appl. Phys. Lett. 88, 031911-3, 2006.
[9] P. Jood, R. J. Mehta, Y. Zhang et al., Al-doped zinc oxide nanocomposites with enhanced thermoelectric properties, Nano Letters, 11, 4337-4342, 2011.
[10] V. Bhosle, J. T. Prater, F. Yang, D. Burk, S. R. Forrest, J. Narayan, Gallium-doped zinc oxide films as transparent electrodes for organic solar cell applications, J. Appl. Phys, 102, 1-5, 2007.
[11] P. S. Venkatesh, V. Ramakrishnan, and K. Jeganathan, vertically aligned indium doped zinc oxide nanorods for the application of nanostructured anodes by radio frequency magnetron sputtering, Cryst. Eng. Comm, 14, 3907–3914, 2012.
[12] J. M. Lin, Y. Z. Zhang, Z. Z. Ye et al. Nb-doped ZnO transparent conducting films fabricated by pulsed laser deposition, Appl. Surf. Sci, 255, 6460–6463, 2009.
[13] C. R. Kumar, R. Subalakshmi, J. Surface Sci. Technol, 31, 176–178, 2015.
[14] A. Verma, F. Khan, D. Kar, B. C. Chakravarty, S. N. Singh, M. Husain, Sol–gel derived aluminum doped zinc oxide for application as anti-reflection coating in terrestrial silicon solar cells, Thin Sol. Films, 518, 2649-2653, 2010.

[15] H. Serier, M. Gaudon and M. Menetrier. Al-doped ZnO powdered materials, Al Solubility limit and IR absorption properties. Solid State Sci, 11 1192 -1197, 2009.
[16] W. H. Zhang, W. D. Zhang and J. F. Zhou, Solvent thermal synthesis and gas-sensing properties of Fe-doped ZnO J. Mater. Sci, 45, 209-241, 2009.
[17] J. J. Kim, J. Y. Bak, J. H. Lee, H. S. Kim, N. W. Jang, Y. Yun. Characteristics of laser-annealed ZnO thin film transistors. Thin Solid Films, 518, 3022–5, 2010.
[18] M. A. Khan, M. S. Kumar, M. N. Khan, M. Ahamed, A.S. Al Dwayyan, Microstructure and blue shift in optical band gap of Nanocrystalline Alx Zn1-xO thin films. J. Lumin, 155 275-281, 2014.
[19] Y. L. Ji, Y. J. Jiang, Increasing the electrical conductivity of poly (vinylidene fluoride) by KrF excimer laser irradiation Appl. Phys. Lett, 89, 221103-3, 2006.
[20] L. Chang, Y. J. Jiang, L. Ji, Improvement of the electrical and ferro-magnetic properties in La 0.67Ca 0.33 MnO3 thin film irradiated by CO2 laser. Appl. Phys. Lett, 90, 082505, 2007.
[21] O.V. Overschelde, G. Guisbiers, M. Wantelet. Nanocrystallization of anatase or rutile TiO_2 by laser treatment, J. Phys. Chem. C, 113 (2009) 15343 -15345.
[22] Q. Xu, R.D Hong, H.L Huang, Z.F Zhang, M.K Zhang, X.P Chen, Zh.Y Wu, Laser annealing effect on optical and electrical properties of Al doped ZnO films, Optics & Laser Technology 45, 513–517, 2013.
[23] L. Hui, T. Yaoquan, X. Lin, B. Fang, D. Luo, A. Laaksonen, Effects of laser irradiation on the structure and optical properties of ZnO thin films. Mater. Lett, 64, 2072-2075, 2010.
[24] F. K. Shan, Y. S. Yu, Band gap energy of pure and Al-doped ZnO thin films. J. Eur. Ceram. Soc, 24, 1869-1872, 2004.
[25] T. Aohi, Y. Hatanka, D. C Look, ZnO Diode Fabricated by Excimer-Laser Doping. Appl. Phys. Lett, 76, 3257-2, 2000.
[26] X. C. Wang, G. C. Lim, W. Liu, C. B. Soh, S. J. Chua, Effects of 248 nm excimer laser irradiation on the properties of Mg-doped GaN. Appl. Surf. Sci, 252, 2071-2077, 2005.
[27] Y. Zhao, Y. Jiang, Effect of KrF excimer laser irradiation on the properties of ZnO thin films J. Appl. Phys, 103, 114903-3, 2008.
[28] M. Nie, Y. Zhao, Y. Zeng, Effects of annealing and laser irradiation on optical and electrical properties of ZnO thin films, J. Laser Appl, 26, 022005-6, 2014.
[29] M. S. Oh, S. H. Kim, D. K. Hwang, S. J. Park, T. Y. Seong, Formation of low resistance nonalloyed Ti/Au ohmic contacts to n-Type ZnO by KrF excimer laser irradiation. Electro-Chem. Solid-State Lett, 8, G317-G319, 2005.
[30] M. S. Oh, D. K. Hwang, J. H. Lim, Y. S. Choi, S. J. Park, Improvement of Pt Schottky contacts to n-type ZnO by KrF excimer laser irradiation. Appl. Phys. Lett, 91, 042109-3, 2007.
[31] Q. Xu, R. D. Hong, H. L Huang, Z. F. Zhang, M. K. Zhang, X.P. Chen, Z.Y. Wu, Laser annealing effect on optical and electrical properties of Al doped ZnO films. Opt. & Laser Technol, 45, 513-517, 2013.
[32] Y. Hou, A. H. Jayatissa, Effect of laser irradiation on gas sensing properties of sol–gel derived nanocrystalline Al-doped ZnO thin films. Thin Sol. Films, 562 585–59, 2014.
[33] D. Scorticati, A. Illiberi, T. Bor, S. Eijt, H. Sch€ut, G. R€omer, M. K. Gunnewiek, A. Lenferink, B. Kniknie, , R. M. Joy, M. Dorenkamper, D. de Lange, C. Otto, D. Borsa, W. Soppe, A. H. in't Veld, Thermal Annealing Using Ultra-Short Laser Pulses to Improve the Electrical Properties of Al:ZnO. Acta Mater, 98, 327-335, 2015.
[34] M. N. Khan, A. Almohammedi, Evaluation of modification in nano structured ZnO film irradiated with Pico second laser at 355 nm wavelength., Sylwan Journal, 163(11), 287-300, 2019.

[35] N. Srinatha, Y. S. No, V. B. Kamble, S. Chakravarty, N. Suriyamurthy, B. Angadi, A. M. Umarjif, W. K. Choib, Effect of RF power on the structural, optical and gas sensing properties of RF-sputtered Al doped ZnO thin films, RSC Adv. 6, 9779, 2016.
[36] A. Djelloul, M. S. Aida, J. Bougdira, Photoluminescence, FTIR and X-ray diffraction studies on undoped and Al-doped ZnO thin films grown on polycrystalline α-alumina substrates by ultrasonic spray pyrolysis, J. Lumin. 130, 2113, 2010.
[37] F. L. Tang, M. K. Wong, J. C. Fung, C. S. Chang, S. T. Lee, Transparent conducting aluminum doped zinc oxide thin film prepared by sol-gel process followed by laser irradiation treatment. 517, 891-895, 2008.
[38] M. J. Akhtar, H. A. Alhadlaq, A. Alshamsan, M. A. M. Khan, M. Ahamed, Aluminium doping tunes band gap energy level as well as oxidative stress mediated cytotoxicity of ZnO particles in MCF-7 cells, Scientific reports, 5, 13876, 2015.
[39] C. Manoharan, G. Pavithra, M. Bououdina, S. Dhanapandian, P. Dhamodharan, Characterization and study of antibacterial activity of spray parolysed ZnO:Al thin films, Appl. Nanosci. 6, 815-825, 2015.
[40] J. Zhang, L. Sun, J. Yin, H. Su, C. Liao, C. Yan, Control of ZnO Morphology via a Simple Solution Route. Chem. Mater. 14, 4172, 2002.
[41] M.H. Huang, Y.Y. Wu, H.N. Feick, N. Tran, E. Weber, P.D. Yang, Catalytic Growth of Zinc Oxide Nanowires by Vapor Transport. Adv. Mater. 13, 13, 2001.

Predicting the Stagnation Time of Covid-19 Pandemic Using Bass Diffusion and Mini-Batch Gradient Descent Models

Mohamed Atef Mosa,
Department of Information Technology, Institute of Public Administration, Riyadh, KSA
Egyptian Space Agency, Assembly, Integration & Testing Center, Cairo, Egypt.
mosamo@ipa.edu.sa, mohammedatefmosa@gmail.com

Abstract: The coronavirus (SARS-CoV-2), which first appeared in Wuhan, China, in December of 2019, spread quickly around the world, eventually categorizing it as a global "Epidemic". In early 2020, the SARS-CoV-2 emerging virus had devastating effects on all aspects of daily life, public health, and even the global economy. During that epidemic, much effort had been made to predict the number of confirmed and deaths, and when the epidemic would subside. However, the prediction of epidemic indications (COVID-19) was highly uncertain and different from what happened next. Multiple and rapid virus mutations, and late detection of infection in many cases of people, have made the prediction process complicated and difficult, with some of the proposed models appearing to be largely misleading. In this research paper, we reviewed the analytical and statistical methods to extrapolate the most important data and indicators about the infection (COVID-19) and the rate of confirmed, recovery, and deaths during the past few months in some countries of the world, especially in the Kingdom of Saudi Arabia. On the other hand, we proposed the time for the infection to subside in the Kingdom of Saudi Arabia and some other countries. In the proposed prediction model, the Bass diffusion model was adopted by combining with the mini-batch Gradient descent algorithm to obtain the optimum values for the Bass algorithm parameters. The model was trained on about 85% of the available historical data and tested on the rest of the data. The proposed model indicated that the Kingdom of Saudi Arabia will face an increase in the coming days in terms of the high number of confirmed cases. Moreover, the rate of increase in injuries will decrease over time until it reaches its lowest levels in January of the next year. The model also showed that the curved flattening point for confirmed figures will be at the mentioned month, which is the expected date for the epidemic to recede in Saudi Arabia in the absence of other aftershocks.

Keywords: Data analysis; prediction; COVID-19; Bass diffusion model; mini-batch gradient descent.

توقع زمن الركود لجائحة كورونا (كوفيد -١٩) باستخدام نماذج انتشار باس ونماذج نزول التدرج ذات الدفعة المصغرة

الملخص: انتشر الفيروس التاجي، فيروس كرونا المستجد (SARS-CoV-2)، والذي ظهر لأول مرة في مدينة ووهان الصينية في ديسمبر من العام ٢٠١٩، لكن سرعان ما توغل الي جميع أنحاء العالم ليصنف في نهاية الأمر على أنه "جائحة" (Epidemic) عالمية. في مطلع العام ٢٠٢٠، كان لفيروس كرونا المستجد SARS-CoV-2 تأثيرات مدمرة على كل جوانب الحياة اليومية والصحية، بل وحتى الاقتصاد العالمي. خلال تلك الجائحة بُذلت جهود كثيرة للتنبؤ بأعداد الإصابات والوفيات وموعد انحسار الوباء. ومع ذلك، فإن التنبؤ بمؤشرات عدوي (COVID-19) كانت غير مؤكدة بدرجة كبيرة ومغايرة لما حدث بعد ذلك. طفرات الفيروس المتعددة والسريعة، والاكتشافات المتأخرة للعدوي في حالات كثيرة، جعلت من عملية التنبؤ أمرا معقدا وعسيرا، بحيث ظهرت بعض النماذج المقترحة على أنها مضللة. في هذه الورقة البحثية استعرضنا بعد الطرق التحليلية والإحصائية لاستقراء أهم البيانات والمؤشرات حول عدوي (COVID-19) ومعدل الإصابات والتعافي والوفيات خلال الأشهر القليلة الماضية في بعض دول العالم وبالأخص في المملكة العربية السعودية. من ناحية أخري قمنا بتسليط الضوء على إمكانية التنبؤ بمعدلات انتشار عدوي (COVID-19) في المملكة العربية السعودية وبعض الدول الأخرى في الأيام القادمة والتنبؤ بموعد انحسار هذا الوباء. تم الاعتماد في نموذج التنبؤ المقترح على خوارزم باس لقياس الانتشار (Bass diffusion model) وذلك بالاندماج مع خوارزم الانحدار التدريجي (mini-batch gradient descent) للحصول على القيم المثلي لمعاملات خوارزم باس. تم تدريب النموذج على حوالي ٨٥٪ من البيانات التاريخية المتاحة حول عدوى (COVID-19)، واختباره على بقية البيانات. أوضح النموذج المقترح أن المملكة العربية السعودية سوف تواجه تزايدا في الأيام القادمة من حيث ارتفاع حصيلة أعداد الإصابات. وأن معدل الزيادة في الإصابات سينخفض مع الوقت حتى يصل الي معدلاته الدنيا في شهر يناير على أقل تقدير من العام القادم وذلك في حالة عدم ارتداد الموجه. كما أظهر النموذج أن نقطة تسطيح المنحني لأعداد الإصابات ستكون خلال الشهر ذاته، وهو الموعد المتوقع لانحسار الوباء بالمملكة العربية السعودية وذلك في حالة عدم وجود موجات ارتدادية أخرى.

1. Introduction:

Coronaviruses are a wide group of viruses that can cause many human losses, ranging from a common cold to severe acute respiratory syndrome. Also, viruses from this group cause many different animal diseases. Since many early cases were associated with a large market for marine and animal food in Wuhan, China, the virus is believed to have an animal origin, but so far it has not been confirmed [1].

Since the emergence of the emerging SARON virus (SARS-CoV-2) in December last year 2019, researchers around the world have developed various models that rely on the data monitored about (COVID-19) infection to predict the damage and injuries that will be left shortly around the countries of the world. Some models have tried hard to find a date for the outbreak of the virus globally or even in certain regions. However, this type of prediction, which is "time-related prediction", may develop some emerging factors that may make it a very accurate prediction. This is because the values of the observed readings change periodically and around the clock. Among the efforts made was a model developed by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington [2], which is updated around the clock. Another model was developed by the MRC Center for Analysis of Global Infectious Diseases in London [3]. Also, some models focused on predicting future mortality numbers and numbers that will need intensive care [4-6]. While many other studies focused on answering two main questions: 1) How many expected cases of infection will be 2) and what is the maximum number of injuries that can be reached [7-9]. On the other hand, the extent of the impact of social separation, travel restrictions, mitigation strategies, and embargoes imposed by governments on societies has been highlighted, to study the impact of these measures on the expected numbers and statistics [8-11]. Some studies published [4-7] also attempted to verify the accuracy of some of the proposed prediction models. However, it was found that the cited models that were published by the IHME had errors and significant deviations from the real numbers [6, 12]. Where it was found that the rate of diffraction in the numbers of real deaths is very large and completely different from the expected numbers from the proposed model [13]. The IHME team later revised the model [5], but prediction errors are still high and out of real numbers. Over time, researchers learn and develop their prediction methods and algorithms to reach near-accurate or largely unrealistic numbers. Despite the intrinsic complexity and uncertainty of predictions of a COVID-19 infection, some of the work

presented significantly affected some of the precautionary policies and procedures in some respects [14, 15].

Uncertainty and ambiguity about the COVID-19 infection increase the desire of countries and governments to know the future expectations and effects that the infection may have, especially those whose economies have been greatly worsened by the pandemic. At the same time, it becomes clear that it is also difficult to complete the prediction accurately due to the mystery surrounding the pandemic. The primary challenge rooted in the disease (COVID-19) as a "wicked problem", which was formulated by researchers Rittel and Webber [16]. Where troublesome problems are described as those new, unique, complex, and evolving problems with incomplete, contradictory, and changing requirements, often difficult to identify. Here are some characteristics of troublesome problems: 1) They are often closely related to ethical, political, economic, or professional issues. 2) It cannot be solved by traditional analytical method. 3) No solution can be objectively tested or evaluated and then confirmed as right or wrong. 4) It makes no sense to talk about optimal solutions to these kinds of problems. 5) The proposal designed to solve the problem may lead to the emergence of other hidden problems.

COVID-19 infection is one of these bothersome problems, which are highly mysterious and not naturally predictable in the general sense. However, this does not mean that objective science-based analyzes and forecasts are completely futile. This only means that the stereotypes of traditional solutions to optimization, accuracy in modeling, and predictions about this pandemic should be avoided. For example, in some of the proposed prediction models for COVID-19 infection [4, 5, 12, 13], there are clear intentions or goals to improve prediction models based on the variance between real and predicted values, but the answer will still be inaccurate and not final to some extent. So talking about the timing of recession may be meaningless and may create a false sense of certainty that is not present. Moreover, when real-world scenarios change in terms of government interventions and human behaviors, it will be naive to assess the accuracy of a model that has been trained using data collected in conditions different from the one it is being tested on and that has been created under completely different conditions. Therefore, to ensure prediction more closely to realism in light of these procedures and changing decisions periodically, different prediction models must be fed with recent data on infection to reduce the error gap that may be present.

Consequently, my infection (COVID-19) requires an innovative strategy to elicit insights based on periodically updated data. This is called predictive monitoring: it is the monitoring of

variable forecasts that are constantly updated with the latest data, along with real data monitoring. Dealing with time-bound predictive monitoring (COVID-19) is not clear as mentioned above, due to several other additional factors: 1) Historical data on the infection (COVID-19) that have been collected, may not be sufficient to allow for a prediction long term. 2) Results may differ according to the precautionary measures emerging by the governments, which may change periodically and around the clock. Add to this the general awareness of the people and their commitment to these measures. For example, Countries that have taken harsh and strict measures against this pandemic, as China has done, have had different results than other countries that initially tolerated precautionary measures, or who recently discovered it. 3) Rapid mutations created by the virus to disguise and adapt to different environments by copying new copies that are difficult to identify. Analysis of more than 5,300 genomes of the Coronavirus in 62 countries showed that although the virus was somewhat stable, some "genomes" were gaining mutations, including a mutation in "Spike protein," the protein the virus uses to infect human cells. Scientists say the genetic change in "Spike protein" is a sign of the virus adapting to its human host. According to researchers at the London School of Hygiene and Tropical Medicine, it is unclear how mutations affect the virus, but these mutations that have arisen independently in different countries may help the virus to spread more easily. 4) Possible bouncing wave scenarios that might strike again and forcefully. As part of the analysts' believes that China did not completely control the virus. As Dr. Li Languan, who led a medical team to combat the virus in Wuhan, said that there are still many patients in critical condition, she also indicated that it could lead to the opening of the Chinese borders with other countries to a second wave of the spread of the virus in China. According to a study conducted by a hospital in Wuhan, it was revealed that 3-14% of those recovering from the virus contract it again without showing any symptoms, and they are called "Silent Carrier". In the same context, "Wang Wei" - director of the "Tongji" Hospital in Wuhan - told CCTV channel that 5 out of 147 recoverees are infected with the virus again. This scenario also reinforces some historical facts, such as the influenza pandemic that spread in 1889 and 1918 and spread in three waves, and each wave was stronger than the one that preceded it as a result of the mutating of the virus. Therefore, if these factors were previously investigated, it would affect the accuracy of the prediction models.

This research paper aims to analyze the data of the COVID-19 pandemic worldwide, especially in the Kingdom of Saudi Arabia, and compare it with some countries to extract some important information from it. It is known that the Kingdom of Saudi Arabia is special, therefore, because

it incubates dozens of different nationalities from all over the world. So, its steps should be more careful than other Middle Eastern countries due to their different nature. Besides, this paper proposes a model for predicting the numbers of casualties expected in the coming days and when the epidemic will subside. The Bass diffusion model was used in combination with the mini-batch Gradient descent model to obtain optimal values for the Bass algorithm coefficients. Predictive monitoring can guide decision-makers in their vision, plans, and future actions to shape a safer and more interactive future for a pandemic.

2. Methodology of work

The work methodology of this paper is to summarize infection data analysis (COVID-19) using the most famous Python language analysis and display libraries (NumPy, Pandas, and Matplotlib). The data was described at the beginning, then the pandemic data was reviewed and analyzed at the level of the countries of the world combined. After that, the analysis was carried out at the level of the Kingdom of Saudi Arabia and compared to some other countries and some neighboring countries to extract some important information and indications from it. In the end, a model based on time-related prediction was built on the data on (COVID-19) infection, which included determining the number of cases of confirmed, death and recovery expected to be reached and determining when the epidemic will recede.

2.1. Data description

Johns Hopkins University has prepared a smart, dynamic (Dashboard) report that is updated around the clock using affected case data. Johns Hopkins University relied on data collected from the World Health Organization (WHO) website and the CDC website for infection (COVID-19). This dataset contains information on the number of cases and deaths and the number of people recovering from (new COVID-19) disease for the year 2019. The data have the following component:

- History of notes (data)
- State or state
- Country / Territory - Observation Country
- Latest update - time UTC
- The daily and cumulative number of confirmed cases as of that date
- The daily and cumulative number of deaths as of that date
- Daily and cumulative number of cases recovered to that date:

2.2. Covid-19 data analysis in world

In this paragraph, some basic information about the infection that was monitored until 22/9/2020 is shown. Table 1 shows the number of countries that have been affected by the epidemic so far, and which exceeded two hundred countries. The table also shows the total number of confirmed cases, which have approached thirteen million, cases of disease recovery, deaths, and active cases, so far, cumulatively. From Table 1, it is possible to deduce the number of active cases, where the active cases are interpreted as the number of confirmed cases - the number of cases recovered - the number of deaths.

Table 1 also shows the number of active cases, which are cases whose condition has not changed from either injury to recovery or death until now. It may be possible that the increase in the number of active cases is a dependable indicator of the numbers of recovering cases or numbers of deaths compared to the number of confirmed cases, as the proportion between them is inverse. In Figure 1, the number of injuries and deaths, and the numbers of recovering cases are shown daily. On the other hand, Figure (2) shows the number of closed cases. Where closed cases mean, they are cases that have been decided and changed from being active cases to cases that have been recovered or cases of death. The number of closed cases is calculated as = the number of recovered cases + the number of deaths. The report notes that closed cases are increasing markedly with higher recovery rates and lower death rates.

Table 1: Basic information about a COVID-19 infection

Parameter	Value
The total number of countries where the disease is spread:	223
The total number of confirmed cases worldwide:	31779835
The total number of cases that have recovered around the world	21890442
Total number of deaths worldwide:	975104
The total number of active cases worldwide:	8914289
The total number of closed cases worldwide:	22865546
The approximate number of cases confirmed daily around the world:	129186
The approximate number of cases recovering per day around the world:	88986
The approximate number of deaths per day worldwide:	3964
The approximate number of confirmed cases per hour worldwide:	5383
The approximate number of cases recovering per hour around the world:	3708
The approximate number of deaths per hour worldwide:	165

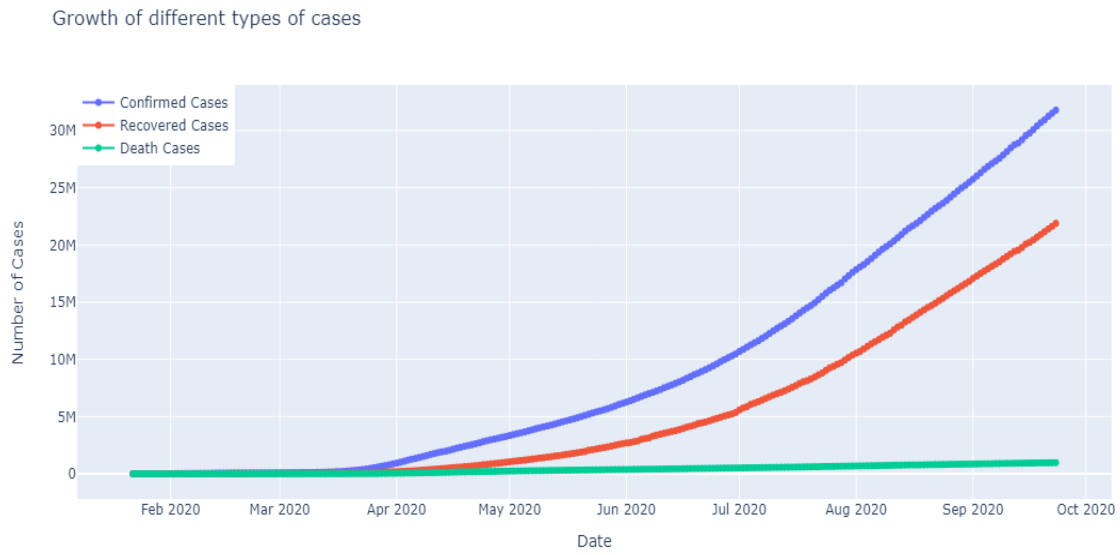


Figure 1: Numbers of infected, recovered, and deceased cases about COVID-19 infection

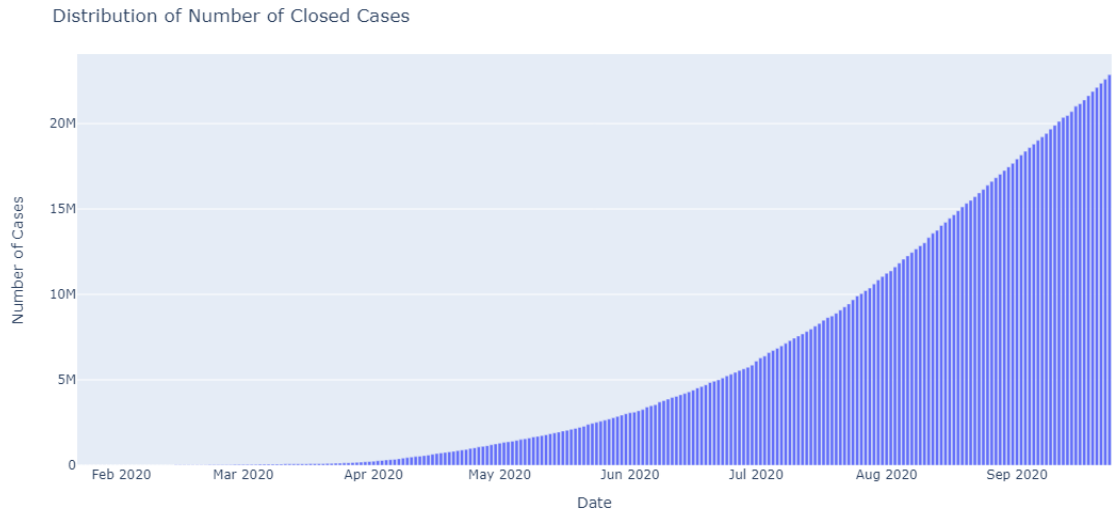


Figure 2: Number of closed cases around COVID-19 pandemic

2.3. Analysis of the Kingdom of Saudi Arabia data

In this section, some basic information about the numbers of COVID-19 infections in Saudi Arabia is presented. Table 2 shows the cumulative number of confirmed and recovered cases, deaths, and active cases as of 22/9/ 2020. Also shown in the table are some statistics of the mean numbers of approximate daily numbers and hourly numbers.

Table 2: Basic information about the Kingdom of Saudi Arabia COVID-19 infection

Parameter	Value
The total number of confirmed cases:	331359
The total number of cases recovered	313786
The total number of deaths:	4569
Total number of active cases	13004
Total number of closed cases	318355
The approximate number of cases confirmed daily:	1609
The approximate number of cases recovering daily:	1523
The approximate number of deaths per day:	22

From Table (2), it turns out that the total number of cases in the Kingdom of Saudi Arabia reached 331359 as of 22/9/2020, while the numbers recovered exceeded 313 thousand, or 94.6%. Looking at the data, we notice that the recovery rate is significantly high and promising. While the data showed that the number of deaths due to the pandemic in the Kingdom reached 4569 cases or 1.3% of the total cases. It is somewhat too low for many countries. The table also shows the approximate average averages for each day and for each hour .

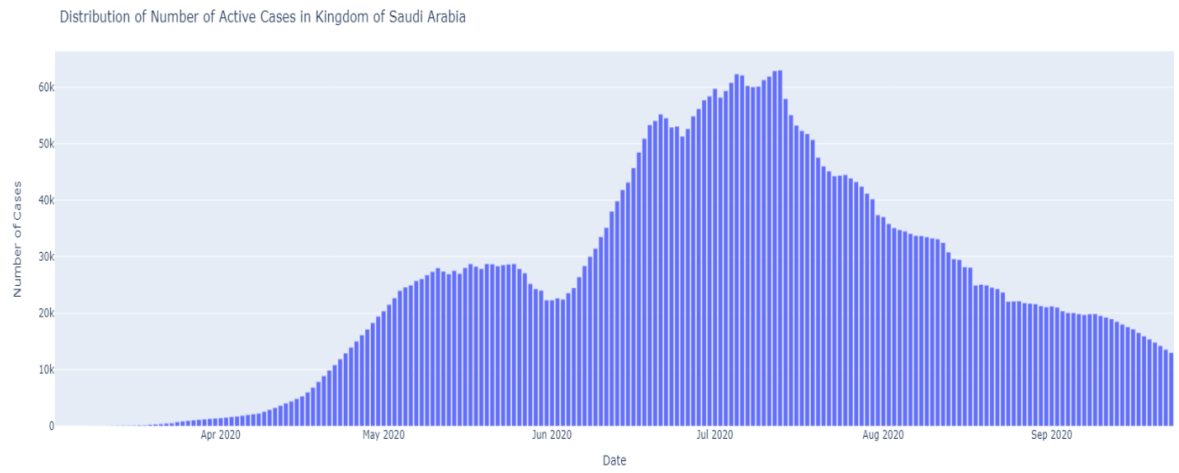


Figure 3: Distribution of numbers of active cases in the Kingdom of Saudi Arabia

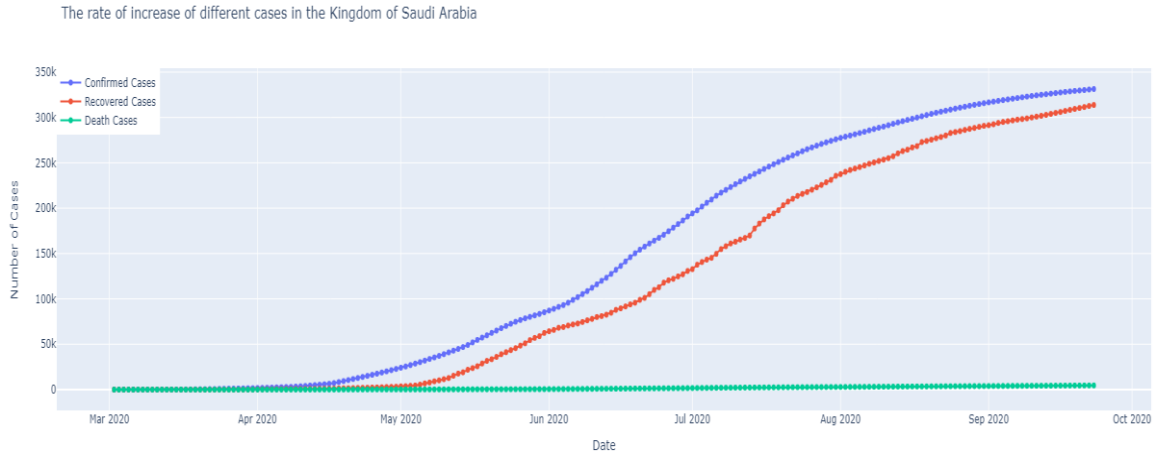


Figure 4: Daily Increase Numbers on the COVID-19 Pandemic Cases in the Kingdom of Saudi Arabia

In Figure (3), the data showed that the rate of active cases was increasing daily, until July 2020, and showed relatively stable stability, until it started to decline at the end of July till now. Figure (4) shows an increase in the number of cases of injury, recoveries, and deceased cases in the Kingdom of Saudi Arabia. The figure shows that at the end of May, the rate of casualty monitoring started to decline significantly and started in early June to the end of July.

3. Prediction

3.1. Bass diffusion model

Predicting the prevalence of COVID-19 is a very complex process. Many algorithms were used for prediction and relied on several techniques including time series [18], support vector machine [19], neural network [20], and other machine learning prediction models [21], but nothing is superior to the Bass diffusion model [22 - 24] is intended to predict prevalence. The Bass model is characterized by its simplicity and accuracy in predicting the average number of consumers for products related to time. There are thousands of academic articles on the Bass model and its applications. Frank M [22] papers one of the most important articles that have cited several citations to be the highest ever in predictive research in general. In this research paper, we seek to employ the Bass diffusion model in combination with mini-batch gradient descent to predict the prevalence of a COVID-19 epidemic in the Kingdom of Saudi Arabia and some other countries, by determining the number of infections that are likely to be reached during the coming months and then determining the date of the epidemic recession.

The nature of the Kingdom of Saudi Arabia differs from most of the countries of the world, due to the presence of the two holy mosques in it, and to attract a large number of expatriate workers of various specialties. In theory, each of these people is vulnerable to a COVID-19

infection. So, we'll assume that there is a fixed potential population for infection denoted by the symbol m. Where m represents the population of a country or city. When the epidemic first sweeps, everyone is at risk of infection. And with many infected people, the number of people exposed to the infection thus decreases. (Here we assume that no one can get COVID-19 infection twice and everyone can get infected at most once) to complete the prediction. The number of infections is denoted by the symbol S, and it changes as a function of time t. The number of cumulative injuries is also denoted by the symbol Y and it also changes as a function of time t.

$\boldsymbol{m}$ *The number of people nominated for infection*

$\boldsymbol{S_{(t)}}$ *The number of people candidate for infection per day* $\boldsymbol{t^{th}}$

$\boldsymbol{Y_{(t)}}$ *The number of cumulative persons who are candidates for infection to day* $\boldsymbol{t^{th}}$

After defining these variables, we will have the following simple definition. That is, the number of cumulative infections is equal to the sum of all infected in days t.

$$\boldsymbol{Y_{(t)} = \sum_{\tau=1}^{t} S_{(t)}} \quad (1)$$

We can calculate the number of people who are not yet infected, and who are denoted by the symbol P_((t)) by equation (2). These people are defined as the number of people at risk of infection m-people who were infected Y_((t))

$$\boldsymbol{P_{(t)} = m - Y_{(t)}} \quad (2)$$

The Bass diffusion model was used to predict the number of buyers in stores based on their historical data at a specific time. So, the buyers were categorized into two categories: Immediate Buyers and Counterfeit Buyers. As for the buyers 1) The initiators: they are the most receptive buyers of the new products and who have a desire to acquire the new products when they are issued. While buyers (2) imitators do not have the lead in purchasing new products, but they do not buy it. In the end, buyers buy the product, but some initially buy it, and some buy it after that. The mathematical model of the Bass diffusion model is known as Equation (3), where the model predicts the number of potential buyers of their types at time t.

$$\boldsymbol{S_{(t)} = \rho(0) + \frac{q}{m} Y_{(t)}} \quad (3)$$

Where the initiators are marked with the symbol ρ (0), they were the ones who initially purchased at zero time. Q indicates the number of imitators who purchased at a later time. To adjust the Bass model and reformulate it to fit the problem of predicting the number of pandemic injuries: We assume that at the beginning, there were certainly many unknown injuries whose owners were injured without their knowledge of the pandemic. These

correspond to the numbers of proactive buyers in the procurement prediction model. These individuals who have been exposed to COVID-19 infection because they were unaware of the pandemic, and who did not initially know it, will be symbolized ρ (0) and they are the people who were infected in time zero. Also, these people marked with the symbol ρ (0) are not the only ones at risk. Other uninfected people deal with these infected people and become infected as well. This second type of person, who was aware of the pandemic after announcing it, and despite these people being aware of the dangers of social participation with the injured, they chose to violate the instructions and not to take into account the precautionary social divergence procedures and allow themselves to be infected with this infection. This group of people is denoted by q. It should be noted that the individuals in the aforementioned categories have clear differences in the mechanism of infection, but all of them are ultimately infected.

The Bass diffusion model is shown in its traditional form, shown in equation (3). But the model in its image is not suitable for application directly in our problem. To be compatible with the problem of infection, the model must be reformulated in another appropriate forum. As for predicting the numbers of buyers, it is somewhat different from predicting the numbers of people injured in the event of a pandemic. In purchases, the number of buyers increases, while in our case the number of injured people increases and decreases at the same time. The truth is that the number of injured increases, but at the same time decreases daily. Once a person has been infected, it is classified as infected and counted as infected. However, the outcome of the injury does not keep the same condition as in the buyers' prediction model. The patient may change his condition to two other cases in the end, or to keep his condition infected as is. The patient may turn from injury to death or recovery. But surely the rate of injuries will be greatly affected by these results. Accordingly, the total number of recovery cases and deaths is referred to as closed cases. From the above, we can say mathematically that closed cases o are divided into two parts. Reciprocal cases of death and deaths.

o The number of closed cases - who switched from injury to another

$O_{(t)}$ *The number of closed cases - who have changed from infection to another until today* $\boldsymbol{t^{th}}$

Therefore, equation (3) for the prediction model will be modified to appear in its new form, which is compatible with the prediction of the pandemic numbers, as with equation (4): where the total number of injuries will be affected by cases of recovery and deaths alongside cases with new cases of infections.

$$\boldsymbol{S_{(t)} = \rho(0) + \frac{q}{m}[Y_{(t)} - O_{(t)}]} \qquad (4)$$

3.2. gradient descent

In the fourth equation for the pandemic injury prediction model, the values of the coefficients ρ (0), q, remain unknown, which relates to the number of casualties for people who are ignorant of the world. M is defined as the number of people at risk in the country for which it is predicted. Y_((t)), O_((t)) are defined as the cumulative number of closed and active cases respectively. While the coefficient values ρ (0), q remains unknown. What is required is to determine the optimal values for them in each country to perform a correct and realistic forecasting process. There are many gradient descent algorithms, which differ in how they work to calculate the gradient of a specific goal function. A gradient slope algorithm is an optimization algorithm that is used to reach the minimum coefficients for a given target function. According to Ruder, S. [25], and after making several comparisons between linear regression algorithms, it was found that the most accurate results and the lower number of turns to reach the optimal coefficients are mini-batch gradient descent. The objective here of applying the linear regression algorithm is to determine the optimum values for the coefficients ρ (0), q so that the prediction values for the incidence cases in the previous period for which we have their data are identical and consistent with the real data in that period in the training phase.

The mechanism of work is divided into two parts: a section for training and a section for forecasting. In the training phase, the model is trained on historical data to arrive at the optimal values for the coefficients, which make the real and predictive readings of that period identical. As for the stage of forecasting, it is the stage in which the numbers of injuries and death are predicted in the coming future days, and the date of flattening of the curve is determined, which is the time of receding of the epidemic and the absence of any additional injuries.

4. Implementation and result analysis

In this section, the parameter setting for the Bass-MBGD model is discussed and computational results are presented.

4.1. Simulated instances

In the implementation phase, the work is divided into three phases: 1) the model training phase. 2) prototype testing stage. 3) Prediction stage. As for the first stage, which is the training phase, the aim was to obtain the optimum values for the coefficients $\rho(0)$ and q for each country which make the real readout values approximately equal to the values predicted by the model. The bass model was implemented to predict future readings related to time for each country as

in equation (4). The inputs to the model are the population of each country, which is what we get from Table (3). As for the population of the Kingdom of Saudi Arabia, the numbers of expatriates were added to the number of citizens due to their large number. We obtain the numbers of confirmed, recovery, and mortality for each country from the data described in paragraph (1.1). Before making the prediction and diffusion model, the stepwise regression algorithm assumes values for the coefficients ρ (0), q, and is offset in equation (4), the actual values compared with the inferred values, and the error amount calculated. The process is done in many circuits until the lowest error value is reached. During the training period of the model on the data of each country, the optimum values for the coefficients ρ (0), q, some of which are mentioned in Table (3), were obtained .

Table 3: Some countries' population numbers and transaction values $\rho(0)، q$

Country	Number of populations	$\boldsymbol{\rho(0)}$	$\boldsymbol{q}$
Brazil	16436526	0.011	12.4579
India	2975443	0.013	13.124
South Africa	2045931	0.007	8.7911
US	1900541	0.046	52.8236
Kuwait	1899508	0.011	15.0557
Russia	734909	0.039	89.7219
Qatar	482567	0.016	15.2599
Egypt	396549	0.009	8.4178
Bahrain	357503	0.008	7.4706
Saudi Arabia	356251	0.022	23.1464
United Kingdom	314587	0.047	60.8451
Italy	236313	0.059	48.2137
Spain	231939	0.087	352.029
Bangladesh	194387	0.024	33.2524
France	186878	0.083	406.142
Turkey	160517	0.073	377.597
Iran	137544	0.041	18.8444
Canada	108062	0.044	49.8065
United Arab Emirates	65392	0.028	25.0833

We assumed that no one could be infected with COVID-19 twice and each person could be infected at most once. We know that this assumption is not realistic, but it is more realistic than to assume that everyone will be infected two or three times, or that 40%, for example, of those infected persons will be infected again. This is because the data for that are not available. It is possible to say that this hypothesis is an indicator of the expected minimum number of confirmed cases, not the maximum. As other factors affect the number of confirmed cases, including 1) If a person is infected more than once, the actual number of populations increases

in each country. 2) If there will be a second and third backlash striking the country, the model will give different results certainly. Therefore, it was necessary to set some convenient assumptions on the basis of which we say that this is the minimum expectation.

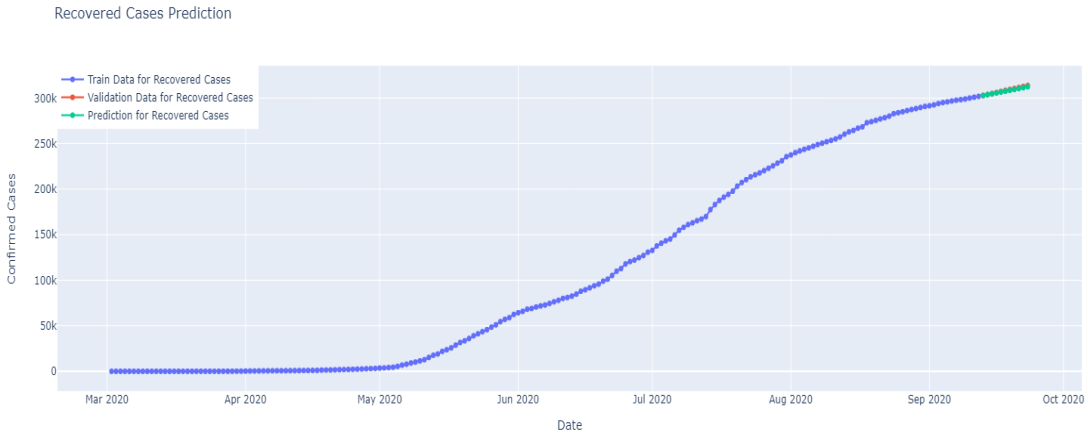


Figure 5: Training and testing period for the model on cases of recovered cases in the Kingdom of Saudi Arabia

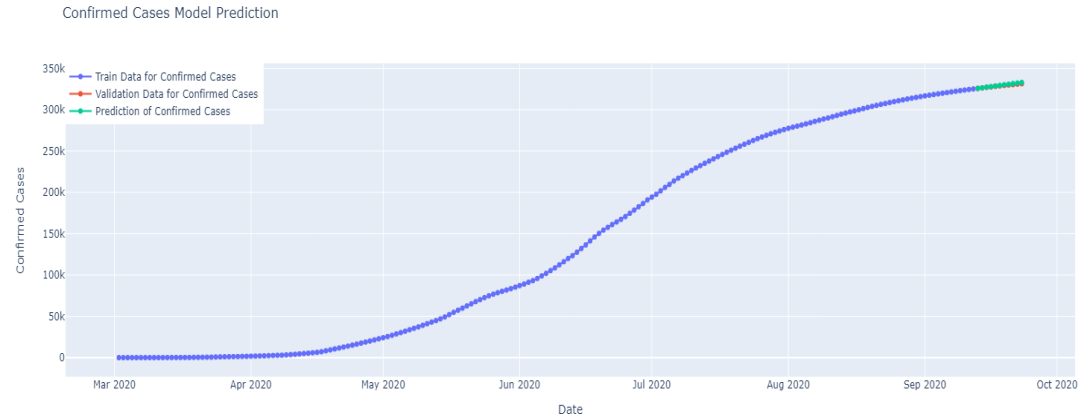


Figure 6: Training and testing period for the model on cases of confirmed cases in the Kingdom of Saudi Arabia

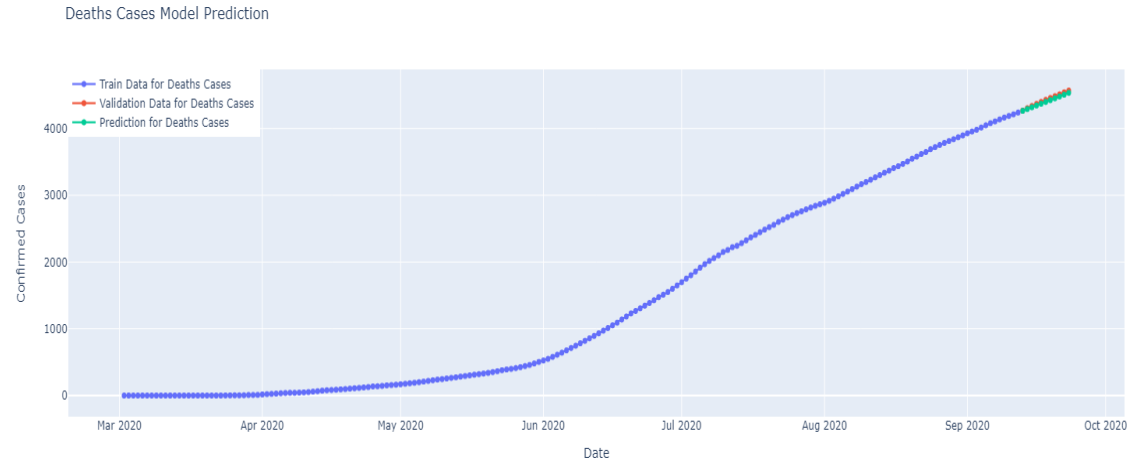


Figure 7: Training and testing period for the model on cases of mortality cases in the Kingdom of Saudi Arabia

4.2. Computational results

The model was trained and tested on numbers for each country separately. where confirmed, active, recovered, and mortality cases will be predicted for the next 500 days. In Figure (5, 6, and 7) the numbers of confirmed, recovered, and mortality cases in the Kingdom of Saudi Arabia are shown as an example. The range of the data in the Kingdom of Saudi Arabia is from the date of the start of monitoring until September 20, 2020. The data is divided into two parts, in a training section the model is represented in about 85% of the data and the rest of the data is used for testing. The blue curve represents the training period of the model in which the optimum values for the parameters of the Bass model for propagation are determined by the gradual regression algorithm. The test period appears later in both the green and red curves, in which the Bass algorithm is applied with the values of the coefficients inferred to predict recovered, confirmed, and death numbers represented by the green line. The readings obtained from the model in the green line appear in the testing phase to a large extent actual data represented in the red line. Note that the model was able to predict the data for this period properly, almost close to the actual readings.

To assess the effectiveness of the Bass diffusion model, we also compared the model in terms of root mean square error (RMSE) with other forecasting algorithms like ARIMA (2,1,1) [26] and the Bass model [27] [28]. In this work, [26] the authors employed the Autoregressive Integrated Moving Average (ARIMA) model to forecast the expected daily number of COVID-19 cases in Saudi Arabia for four weeks. Bass Model [27] is developed for deaths for the period, March 21 to April 30 for the USA as a whole and as the US States of New York, California, and West Virginia. And finally, in this study, [28] the authors adapted the bass diffusion model to determine the time when the COVID-19 curve flattens in the Philippines. Further, it also determined the possible incidence of the second wave of infection.

Table 4: ARIMA order selection based on AIC and BIC approaches in Saudi Arabia for confirmed cases

	AIC	BIC
ARIMA (2,1,1)	3133.5	3144.7
ARIMA (2,1,2)	3134.2	3141.5
ARIMA (1,2,1)	3110.8	3119.8
ARIMA (1,2,2)	3101.9	3111.9
ARIMA (2,2,2)	3138.2	3149.6

All of the comparison algorithms have been re-implemented and experimented on the same data. Moreover, Akaike information criterion (AIC), and Bayesian information criterion (BIC)

criterion are employed for model selection among a finite set of models ARIMA in Kingdom of Saudi Arabia for confirmed cases. As we see in table 4. The best model of ARIMA is ARIMA (1, 2, 2). On the other hand, we compared all models in terms of Root Mean Square Error RMSE. As presented in table 5, based on our results, the prediction methods of our model performed better than other models. According to table 5. we can see that the Bass-GD model got the lowest RMSE in the most sample, and therefore it should be able to predict the number of confirmed, recovered, and deaths cases of COVID-19 in Saudi Arabia and other countries in the next coming months better than the other models.

Table 5: RMSE for different models in some countries for confirmed cases

		ARIMA (1, 2, 2) [26]	Bass model [27]	Bass model [28]	Bass-MBGD
Kingdom of Saudi Arabia	Confirmed	435.20	441.19	443.05	428.06
	Recovered	891.59	900.73	901.64	878.05
	Deaths	82.63	30.82	33.06	27.30
Russia	Confirmed	855.94	8731.46	8806.50	8560.80
	Recovered	20169.92	20308.41	20487.75	19926.80
	Deaths	767.30	748.55	776.31	750.30
UAE	Confirmed	149.13	146.14	149.44	144.40
	Recovered	300.11	302.16	304.79	292.60
	Deaths	13.30	16.36	13.43	8.20
Egypt	Confirmed	129.77	124.62	133.72	122.30
	Recovered	224.03	232.57	228.54	219.40
	Deaths	99.90	108.02	112.93	101.10
United Kingdom	Confirmed	1279.27	1322.29	1330.89	1288.80
	Recovered	2657.67	2623.20	2699.62	2622.20
	Deaths	116.33	109.10	121.09	110.01

After training the model and adjusting the values of the transactions, the data of five hundred days in the future was extrapolated through the last stage, the stage of predicting the numbers of confirmed, recovered, active, and mortality for several countries during the coming months, as shown in Figure (8-12). The dark lines represent the actual numbers of confirmed, deaths, recovered, and active respectively. The faint lines belong to the model. In Figure (8) for the Kingdom of Saudi Arabia, the model predicts that the number of confirmed numbers in the Kingdom will be at least more than 355 thousand cases, in case no second wave. But if the second wave comes, the figures will be increased greatly. and it indicated that the number of confirmed cases will decline significantly during December 2020 until reaching the curve

flattening of injuries in January of the year 2021, which is the expected date for the recession of the pandemic in the Kingdom. The model also showed that the death rate will remain significantly low about the number of injuries in case no second wave. In figures (9-12), expected confirmed, recovered, deaths and active figures for some countries are plotted.

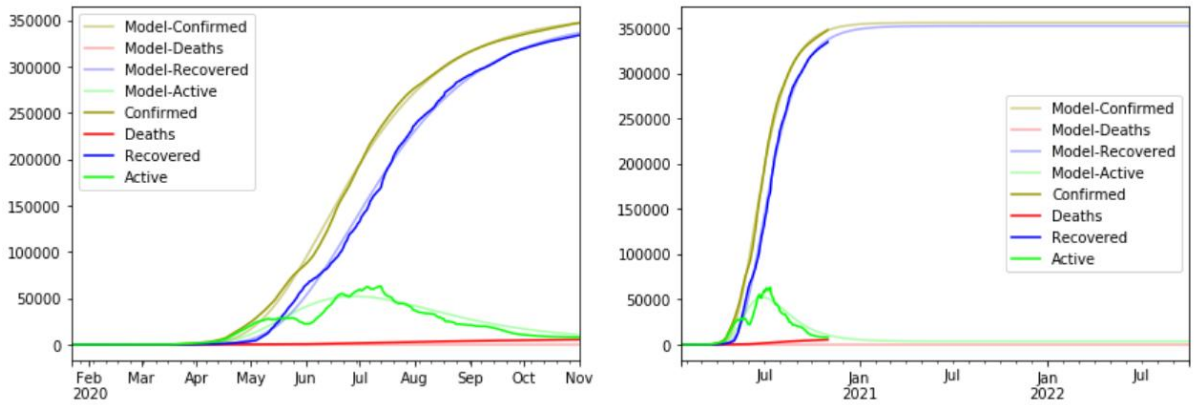


Figure 8: Predicting the number of future cases in the Kingdom of Saudi Arabia

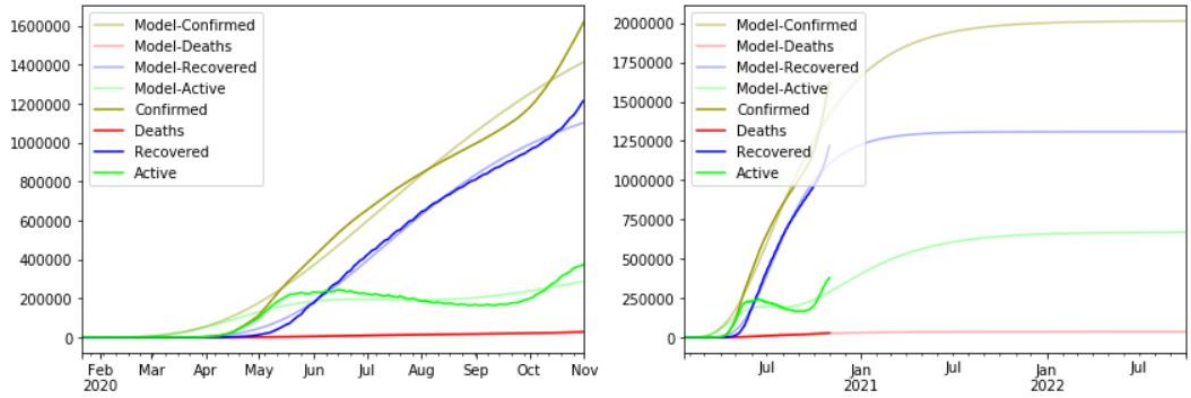


Figure 9: Predicting the number of future cases in the Russia

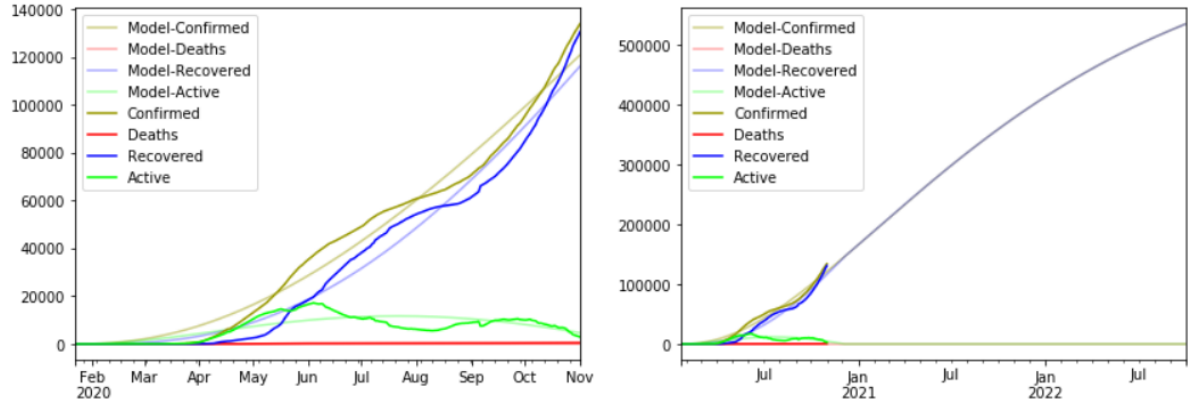


Figure 10: Predicting the number of future cases in the UAE

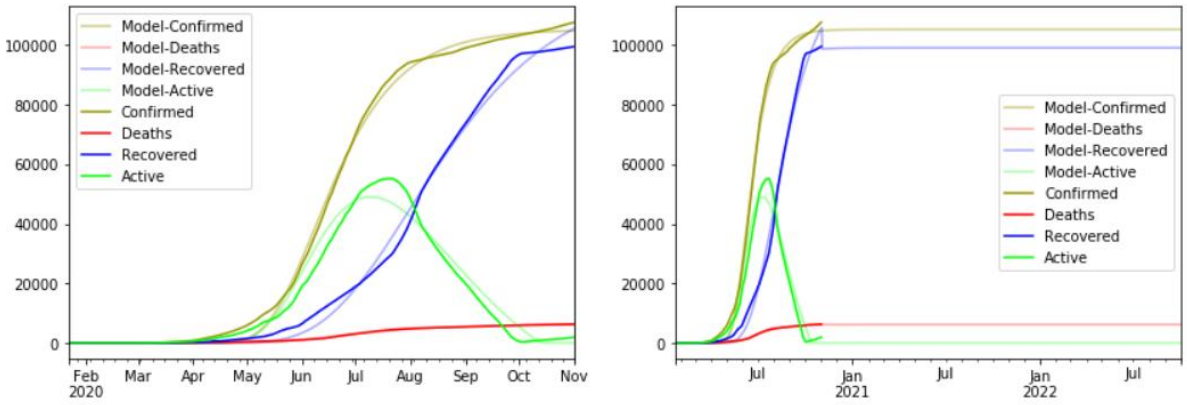


Figure 11: Predicting the number of future cases in Egypt

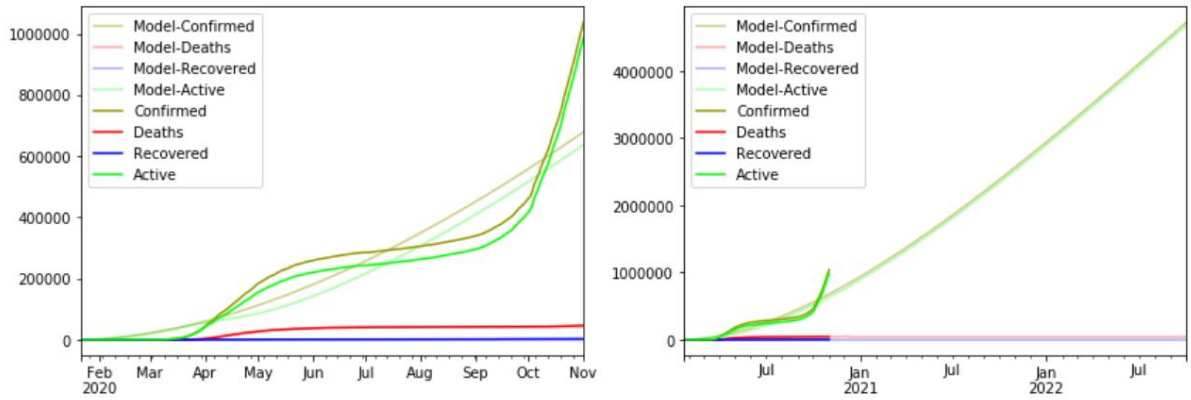


Figure 12: Predicting the number of future cases in the United Kingdom

5. Conclusion and future work:

In this research paper, we reviewed the analytical and statistical methods to extrapolate the most important data and indicators about the COVID-19 infection and the rates of confirmed, recovery, and mortality during the past few months, especially in the Kingdom of Saudi Arabia. On the other hand, the process of predicting the rates of infection prevalence (COVID-19) in the Kingdom of Saudi Arabia and some other countries during the coming days was done by relying on the Bass diffusion model in combination with the mini-batch Gradient descent model. The model was trained on 85% of historical data from the beginning of the actual and significant change around my infection (COVID-19) and tested on the rest of the data. The proposed model showed that the Kingdom of Saudi Arabia will face an increase in the coming days in terms of an increase in the number of confirmed, deaths, recovery, and active cases. And that the rate of increase in injuries will decline over time until it reaches its lowest level in January of next year at the very least, knowing that it is possible that this period will increase

a little. The model also showed that the curved flattening point for the numbers of injuries is likely to be during the month of January 2021, which is the expected date for the epidemic to recede in the Kingdom of Saudi Arabia and to record a rate of zero injuries in the absence of other aftershocks.

In the next few days, the model will be developed to predict other aftershocks, and other models will be developed for the prediction.

6. References

[1] "Update and Interim Guidance on Outbreak of 2019 Novel Coronavirus (2019-nCoV) in Wuhan, China CDC Health Update". New Jersey Department of Health. 18 January 2020

[2] Institute of Health Metrics and Evaluation (IHME) at University of Washington https://covid19.healthdata.org/united-states-of-america

[3] MRC Centre for Global Infectious Disease Analysis at the Imperial College. https://mrc-ide.github.io/covid19-short-term-forecasts/index.html

[4] IHME COVID-19 Health Service Utilization Forecasting Team and Christopher JL Murray. "Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months," 2020. MedRxiv https://doi.org/10.1101/2020.03.27.20043752

[5] IHME COVID-19 health service utilization forecasting team and Christopher JL Murray. "Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries," 2020. MedRxiv https://doi.org/10.1101/2020.04.21.20074732

[6] Woody S., et al. "Projections for first-wave COVID-19 deaths across the US using social-distancing measures derived from mobile phones," 2020. MedRxiv https://doi.org/10.1101/2020.04.16.20068163

[7] Yang Z., et al. "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions." J Thorac Dis, 12(3): 165-174, 2020. DOI: 10.21037/jtd.2020.02.64

[8] Ferguson N.M., et al. "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College London." 16 March 2020. Accessed at www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-03-16-COVID19-Report-9.pdf.

[9] Petropoulos F., Makridakis S. "Forecasting the novel coronavirus COVID-19," PLoS ONE, 15(3): e0231236, 2020. https://doi.org/10.1371/ journal.pone.0231236

[10] Kissler S,M. et al. "Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period." Science, 14 April 2020. DOI: 10.1126/science.abb5793

[11] Chinazzi M., et al. "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak." Science, 24 April 2020, 395-400

[12] Jewell N.P., et al. "Caution Warranted: Using the Institute for Health Metrics and Evaluation Model for Predicting the Course of the COVID-19 Pandemic." Ann Intern Med, 14 April 2020. DOI: https://doi.org/10.7326/M20-1565

[13] Marchant R., et al. "Learning as we go: An examination of the statistical accuracy of covid19 daily death count predictions," 2020. DOI: https://doi.org/10.1101/2020.04.11.20062257

[14] Grey S. and MacAskill A. “Special Report: Johnson listened to his scientists about coronavirus - but they were slow to sound the alarm.” Reuters, April 7, 2020. https://reut.rs/2Rk3sa7
[15] Resnick B. “The White House projects 100,000 to 200,000 Covid-19 deaths.” Vox, March 31, 2020. https://www.vox.com/science-and-health/2020/3/31/21202188/us-deaths-coronavirus-trump-white-house-presser-modeling-100000
[16] Rittel H.W.J. and Webber M.M. "Dilemmas in a General Theory of Planning." Policy Sciences, 4 (2): 155–169, (1973). doi:10.1007/bf01405730
[17] Kermack W. O. and McKendrick A. G. (1927) “A Contribution to the Mathematical Theory of Epidemics,” Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, Vol. 115, 772, pp. 700-721.
[18] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.
[19] Du, X. F., Leung, S. C., Zhang, J. L., & Lai, K. K. (2013). Demand forecasting of perishable farm products using support vector machine. International journal of systems Science, 44(3), 556-567.
[20] Kuo, R. J. (2001). A sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm. European Journal of Operational Research, 129(3), 496-517.
[21] Mahajan, Vijay, Muller E, and Wind,Yoram. (2000) New-Product Diffusion Models, vol. 11, New York: Springer Science & Business Media.
[22] Bass, Frank M. (1969) “A New Product Growth for Model Consumer Durables” Management Science, Vol. 15, 5, pp. 215–227.
[23] Bass, Frank M., Krishnan, Trichy V. and Jain, Dipak C. (1994) "Why the Bass Model Fits without Decision Variables" Marketing Science Vol. 13, 3, pp. 203-223.
[24] Bass, Frank M. (2004) “Comments on ‘A New Product Growth for Model Consumer Durables the Bass Model’” Management Science Vol. 50, 12, pp. 1833–1840.
[25] Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.
[26] Alzahrani, S. I., Aljamaan, I. A., & Al-Fakih, E. A. (2020). Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. Journal of infection and public health, 13(7), 914-919.
[27] Gurumurthy, K., & Mukherjee, A. (2020). The Bass Model: a parsimonious and accurate approach to forecasting mortality caused by COVID-19. International Journal of Pharmaceutical and Healthcare Marketing.
[28] Zeny L. Maureal, Jovelin M. Lapate, Madelaine S. Dumandan, Vanda Kay B. Bicar, Derren N. Gaylo (2020). Adapted Bass Diffusion Model for the Spread of COVID-19 in the Philippines: Implications to Interventions and Flattening the Curve. International Journal of Innovation, Creativity and Change. www.ijicc.net Volume 14, Issue 3, 2020.

بسم الله الرحمن الرحيم والصلاة والسلام على أشرف المرسلين سيدنا محمد وعلى آله وأصحابه اجمعين. أما بعد،

فإنه يسعدني ويشرفني ان أقدم بين أيديكم الاصدار الرابع لسنة ١٤٤٢ هـمن مجلة الجامعة الإسلامية للعلوم التطبيقية. ولقد أنشأت هذه المجلة بهدف أن تجعل من صفحاتها ميداناً لكلِ ما هو جديدٌ ومتميز من الأبحاث العلمية في مجالات الهندسة والعلوم والحاسوب من شتى أنحاء البسيطة.

وتود هيئة التحرير والهيئة الاستشارية أن تتقدم بخالص الشكر لكلِ من ساهم ببحثٍ أو مقالةٍ في هذه المجلة. كما تسرها دعوة الباحثين للمساهمة في الأعداد القادمة بإذن الله.

فبكل فخر واعتزاز اسمحوا لي أن أقدم إليكم العدد الرابع لهذه المجلة وأتمنى لكم قراءة ممتعة.

رئيس هيئة التحرير

د. محمد عبد الرؤوف عابدين

قواعد النشر في المجلة (*)

- أن يكون البحث جديداً؛ لم يسبق نشره.
- أن يتّسم بالأصالة والجدّة والابتكار والإضافة للمعرفة.
- أن لا يكون مستلًا من بحوثٍ سبق نشرها للباحث/للباحثين.
- أن تراعى فيه قواعد البحث العلميّ الأصيل، ومنهجيّته.
- أن يشتمل البحث على:
  - صفحة عنوان البحث باللغة الإنجليزية
  - مستخلص البحث باللغة الإنجليزية
  - صفحة عنوان البحث باللغة العربية
  - مستخلص البحث باللغة العربية
  - مقدمة
  - صلب البحث
  - خاتمة تتضمّن النّتائج والتّوصيات
  - ثبت المصادر والمراجع
  - الملاحق اللازمة (إن وجدت)
- في حال (نشر البحث ورقياً) يمنح الباحث نسخة مجانية واحدة من عدد المجلة الذي نشر بحثه، و (١٠) مستلات من بحثه.
- في حال اعتماد نشر البحث تؤول حقوق نشره كافة للمجلة، ولها أن تعيد نشره ورقيّاً أو إلكترونياً، ويحقّ لها إدراجه في قواعد البيانات المحلّيّة والعالمية- بمقابل أو بدون مقابل- وذلك دون حاجة لإذن الباحث.
- لا يحق للباحث إعادة نشر بحثه المقبول للنشر في المجلة- في أي وعاء من أوعية النشر- إلا بعد إذن كتابي من رئيس هيئة تحرير المجلة.
- نمط التوثيق المعتمد في المجلة هو نمط (IEEE).

(*) يرجع في تفصيل هذه القواعد العامة إلى الموقع الإلكتروني للمجلة: https://jesc.iu.edu.sa

معلومات الإيداع

النسخة الورقية:

تم الإيداع في مكتبة الملك فهد الوطنية برقم ٨٧٤٢ / ١٤٣٩ وتاريخ ١٧ / ٠٩ / ١٤٣٩ هـ

الرقم التسلسلي الدولي للدوريات (ردمد) ٧٩٣٦ - ١٦٥٨

النسخة الإلكترونية:

تم الإيداع في مكتبة الملك فهد الوطنية برقم ٤٢٨٧ / ١٤٣٩ وتاريخ ١٧ / ٠٩ / ١٤٣٩ هـ

الرقم التسلسلي الدولي للدوريات (ردمد) ٧٩٤٤ - ١٦٥٨

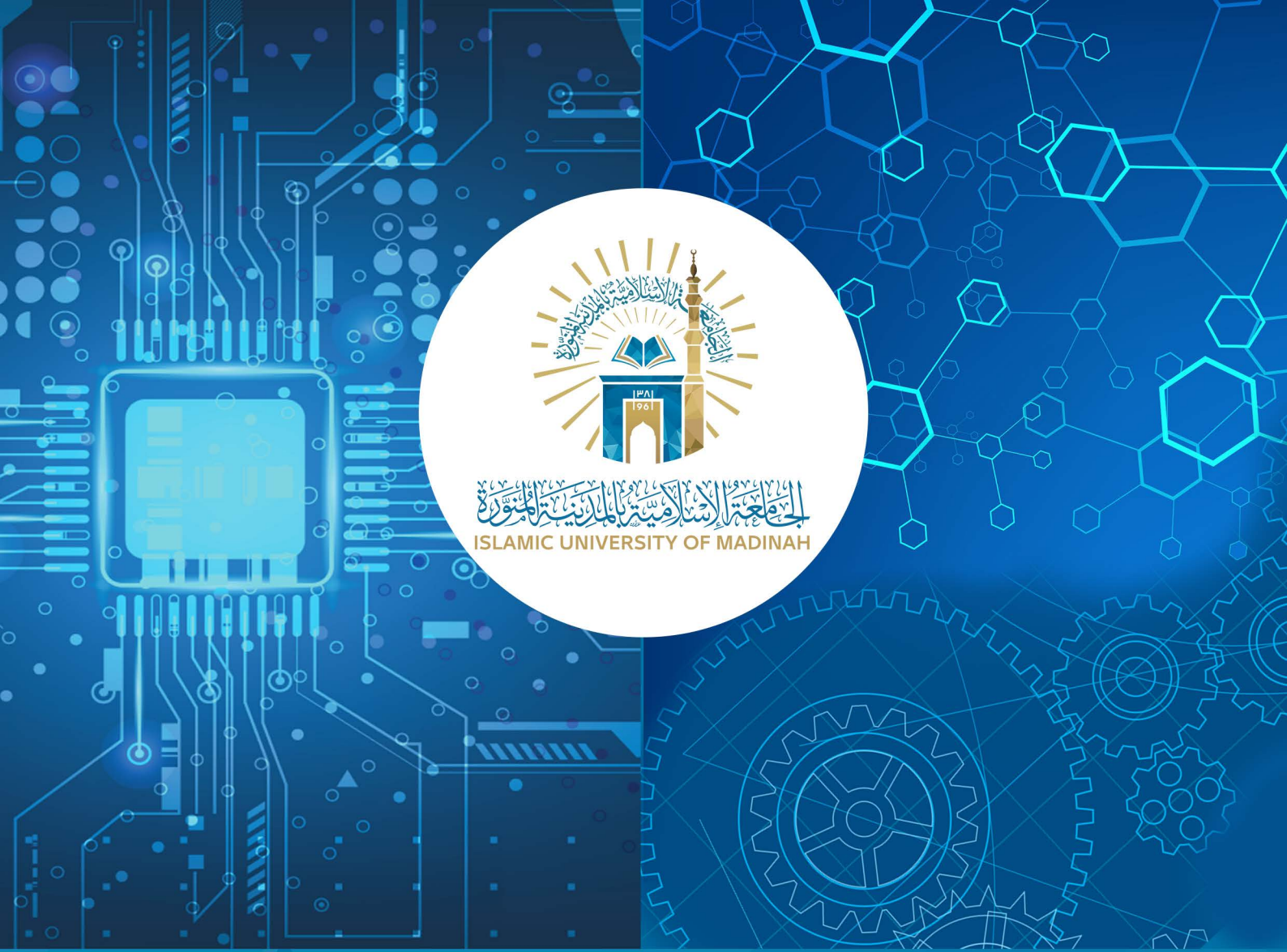
الموقع الإلكتروني للمجلة:

https://jesc.iu.edu.sa

ترسل البحوث باسم رئيس تحرير المجلة إلى البريد الإلكتروني:

jesc@iu.edu.sa


بسم الله الرحمن الرحيم



مجلة الجامعة الإسلامية

للعلوم التطبيقية

مجلة علمية دورية محكمة

العدد: ٢ السنة: الثانية ١٤٤٢هـ