Journal of Natural Sciences and Engineering

Publisher

International Burch University, Sarajevo, Bosnia and Herzegovina

Editorial Office and Administration

International Burch University

Francuske revolucije bb, 71210 Ilidža, Sarajevo, Bosnia and Herzegovina

Website: https://www.ibu.edu.ba/journal-of-natural-sciences-and-engineering/

ISSN 2637-2835 (Print)

Content

Respected readers,

Journal of Natural Sciences and Engineering (JONSAE) is a peer-reviewed biannual journal that aims at the publication and dissemination of original research articles on the latest developments in the fundamental theory, practice and application of engineering, science, and technology.
We provide a platform for researchers, academicians, professionals, practitioners, and students to impart and share knowledge in the form of high-quality empirical and theoretical research papers. The journal covers all areas of Genetics and Bioengineering, Electrical and Electronics Engineering, Information Technology, Architecture, Applied Mathematics, Computer Sciences, and Civil Engineering.

In this issue, I would like to express my gratitude to our authors and manuscript reviewers for their unselfish effort. You have continued your commitment and diligence to the Journal in helping us to produce quality and meaningful content that has the opportunity to advance the field. All of us had personal and professional challenges due to the pandemic and yet despite this, we managed to produce quality content and improve JONSAE's indexation and future perspectives in terms of its metrics and quality growth. Thank you all for being part of this wonderful academic endeavor.

We have both original research and literature review articles published in this issue. Our readers can learn details about how data technology and engineering ventures can contribute to medicine in the form of predictions for depression and anxiety as well as how to use machine learning-based gene clustering in brain cancer classification. Also, you can find out what is the potential of biomaterial-based solutions in cancer research treatment and can homicides be solved using the classification method. In the end, you can read about a staple gene in pharmacogenomics research, *CYP4F2*3*, and the effect of its polymorphism on different traits, including optimum anticoagulation therapy administration based on the patient's genetics.

In the end, we plan to build upon the excellent editorial infrastructure in the next years. In addition to managing the normal turnover of the Editorial Board members, we will seek to expand it by recruiting additional members who can provide expertise in areas that are not currently well represented. In particular, we hope to recruit board members with expertise in such areas as statistical and biostatistical methods, computer science, bioengineering, and biomedical engineering, among others. We also hope to leverage the expertise of the Editorial Board members to train the next generation of scholars and potential Editorial Board members by finding ways to pair student reviewers with senior reviewers for a peer-review mentorship as a part of building a better research environment for new scientists. A similar approach is planned for the administrative members of the Editorial Board, with the aim of improving the quality of the overall Journal design and acquiring a modern visual identity.

Having in mind all stated, I wholeheartedly invite you to read this issue and join our team.

Yours sincerely,

Adna Ašić, PhD
Editor in Chief

The Potential of Biomaterial-Based Solutions in Cancer Research and Treatment

Hannah Abigail Boone*
*International Burch University,
Sarajevo, Bosnia and Herzegovina
hannah.boone@stu.ibu.edu.ba

Literature review

Abstract: *Cancer is a very troubling disease due to its unique morphological characteristics, capacity for drug resistance, and immunosuppressive abilities. Traditional methods used both for research of cancer and its subsequent treatment have fallen short of being able to accurately understand and ultimately defeat cancer within the body. Biomaterials present a unique solution to many problems associated with cancer. The use of biomaterials in cancer cell modeling has promoted a better understanding of tumor microenvironments. Biomaterials can also serve as drug and adjuvant carriers that are more likely to reach their target cancer cells. Many biomaterials also have standalone antitumor properties, and can also help in modulating the immune response, triggering various immune cells to attack cancerous cells. Naturally derived biomaterials include polysaccharides, lipids, polypeptides, vitamin E derivatives, and even plant extracts like curcumin. Biomaterial-based cancer treatments tend to have a longer-lasting and more dependable effect inside the body and can come in many different forms, from polymeric scaffolds to injectable nanoparticles.*

Keywords: Adjuvant therapy, biomaterials, cancer treatment, carrier, immunotherapy.

1. Introduction

One of the most terrifying diagnoses to receive is the diagnosis of having any kind of cancer. Cancer is a very complex disease, as it occurs after mutations in the DNA of the cell cause morphological mutations that fundamentally change the biology and biochemical processes of the cell [1]. Cancer cells have many notable hallmarks, including their ability to divide indefinitely, their tendency to rapidly undergo glycolysis to secrete lactate even in the presence of oxygen, angiogenesis, and decreased cell senescence and apoptosis [1], [2]. To make matters worse, cancer cells can masterfully disguise themselves from immune cells such as cytotoxic T lymphocytes (CTLs) and dendritic cells (DCs), and can even recruit cells such as T regulatory cells (Tregs) and myeloid-derived suppressor cells (MDSCs) to keep other immune cells from recognizing them as cancerous [3], [4], [5]. Additionally, aggregated cancer cells form tumors with complex microenvironments, a phenomenon still not completely understood by scientists [6]. The combination of these factors stipulates that cancer is an incredibly difficult disease to treat, and it is no surprise that it has been considered a death sentence in the past [2].

However, there is hope to both biochemically understand and successfully treat cancer. One stride made in this area is the paradigm shift in how cancer is viewed, as some researchers and pharmaceutical industries are choosing to see cancer as a chronic illness with potentially life-threatening complications, rather than a life-threatening disease in and of itself [2]. Furthermore, toxic chemotherapeutic and radiotherapeutic treatments are no longer the only options for cancer treatment. This is due to the exploration of biomaterials as drug carriers, anticancer therapeutics, and adjuvant treatments [3], [7]. Biomaterials are uniquely suited to treat cancer as they are biocompatible with normal as well as cancerous tissue, allowing them to get close to their target cells, which is something that direct drug injections are often unable to do [2], [8]. Additionally, biomaterials allow for in-depth 3D modeling of cancerous tumors and microenvironments, which has led to a deeper understanding of cancer and how to treat it [9]. This review aims to summarize how biomaterials can be used in both cancer research and treatment, in hopes of raising awareness of alternative solutions to conventional practices that are currently heavily relied upon and may ultimately harm patients.

2. The Cancer Microenvironment Challenge and Using Biomaterials as a Solution

One of the most challenging aspects of laboratory-based cancer research is the accurate modeling of cancer cells and metastasizing tumors. Cancer drug screening experiments have often been conducted using standard 2D cell culture for many years, which has fallen short of delivering a successful picture of how cancer drugs will affect tumors inside an *in vivo* setting [10]. The main limitations of 2D cell culture for cancer drug screening lie in the inadequate physiological and biological complexity of these systems, as well as a shortage of environmental factors that are present in a tumor, such as supporting secondary cell types and extracellular matrix (ECM) proteins [8]. For cancer cells to be properly understood and screened for therapies, a proper tumor microenvironment (SME) must be created [10]. TMEs can be cultured using various biomaterials and tissue engineering techniques [9]. There are two considerations to take into account when studying the TME: the physical and the chemical cues which indicate cancer and its subsequent drug resistance [10].

Physical cues are all indications of a tumor's physical ability to resist cancer drugs. The first cue is just the physical barrier presented by the tumor, which can increase drug resistance by simply keeping active substances away from the tumor core [10]. One of the most important physical cues is the structure of cancer's ECM [9]. The specific ECM around a cluster of cancer cells or tumors can contribute to cancer drug resistance because the ECM creates an environment that isolates the cancer cells from substances that might be harmful to them [6], [11]. ECM produced by cancer cells can undergo a process known as matrix stiffening, where collagen-

modifying enzymes such as proline hydroxylase, lysyl oxidase (LOX), and lysine hydroxylase cause the ECM to remodel itself to become harder; it is often an indicator of cancer progression [2]. The ECM can also secrete substances to make cancer cells more adhesive and can also activate anti-apoptotic signals [6], [10]. Additionally, TMEs inside the ECM produce biochemical cues, which include responses to hypoxia, where they may stop proliferating but can still withstand cytotoxic agents [8]. They can also make different changes with regards to pH, which translates to resistance in the form of ion trapping, efflux pumps, and resistance to acidic pH [10]. Additionally, the ability of tumors to interact with other cells outside the TME can trigger immune-suppressive responses for immune cells and anti-apoptotic responses within cancer cells [4], [10].

Many of these factors involved in TMEs cannot be properly simulated using standard 2D culturing. Fortunately, several methods for cancer cell culturing using biofilm-based 3D techniques exist [9]. Biomaterial-based research for cancer modeling is currently primarily conducted with synthetic polymers, such as polyethyleneimine (PEI) and polyethylene glycol (PEG), which researchers have begun using for the design of "intelligent" biomaterials to simulate the growth conditions of cancer cells, including spheroid formation techniques, 3D bioprinting, and organ-on-chips [2]. Polymeric scaffolds can be used to model the unique ECM environment present around cancer cells discussed previously [10]. Another technique under exploration is injection molding, where a mixture of cells containing some sort of naturally derived or synthetic cross-linking polymer, such as hydrogel or collagen, is injected into a highly detailed and specific mold [2], [8]. Microfluidic cell culture systems, which allow the manipulation of incredibly minuscule amounts of fluid (down to the nanoliter) can also be used to shed insight into cancer biology [10]. Additionally, bioprinting can be used to create specific cancer cell systems for drug testing [2]. Continued research into the use of biomaterials for 3D modeling of cancer cells and their TMEs could even shed light onto the more elusive concepts of cancer growth, such as immune response and microbiome affectation, as well as how cancer cells can alter their surrounding environments to ensure their survival [6].

3. Cancer Treatment Using Naturally Derived Biomaterials

Biomaterial-based cancer treatment can act against tumors and cancer cells in the following ways: by carrying a drug or adjuvant to the site of cancer cells or tumors, by having direct antitumoral properties, or by modulating the immune system and targeting it to attack cancer cells [2]. The reason biomaterials are increasingly explored as potential cancer treatments lies in the abilities of various biomaterials to better penetrate and attach to the TME of a tumor [2]. Biomaterial-based carriers can take the form of 3D structures such as scaffolds and conjugates generally formed by polymers of polysaccharides [12]. Injectable carriers such as liposomes, inorganic nanoparticles, or polypeptides can be injected into the body and then allowed to find their target [3]. While the biophysics and chemistry behind biomaterial carriers are, of course, complex in nature, the concept is fairly simple: a drug, nucleic acid sequence, adjuvant, or other anticancer substance is loaded onto a specific carrier, often labeled with receptors that will allow the carrier to bind to cancer cells [2], [3]. Once bound to the target cells, the carrier releases its contents. Some carriers that have adhesive properties, primarily polysaccharide-based carriers, can remain attached to a cluster of cancer cells, allowing a substance delivered to have a longer effect [2]. Certain biomaterial-based carriers have anticancer properties in and of themselves, such as certain polar lipids, polymers, and polysaccharides, and some have characteristics that make them initially compatible and therefore recognizable by cancer cells, such as in the case of hyaluronic acid [2], [8].

Since cancer cells are exceptional at evading immune response, an important utilization of biomaterials in cancer treatment lies in the realm of immunotherapy. Biomaterial-based immunotherapeutics can have a longer-lasting effect when treating cancer than simple injections

of chemotherapeutic drugs or radiotherapy [12]. There are multiple ways in which biomaterials can stimulate *in vivo* immune responses to cancer cells. Biomaterials containing both antigens and immune adjuvants can stimulate DCs and other antigen-presenting cells (APCs) that may have been inactivated by cancer cells [7], [13]. Another method for stimulating an immune response to cancer is by interfering with T cells whose receptors do not recognize cancer cells as pathogens due to inhibitory factors such as PD-1/PD-L1 and CTLA4 released by the cancer cells; by injecting an anti-PD-1/PD-L2 and CTLA4 antibodies, T cells can be reprogrammed to view cancer cells as targets [7]. Additionally, immune-stimulating cytokines such as interleukin-2 (IL-2) and interferon-alpha (IFN-alpha) can be coupled to a biomaterial carrier and injected to serve as an immune response adjuvant by stimulating the production and differentiation of CDLs [3]. Other adjuvants include toll-like receptor (TLR) agonists, pathogen-associated molecular pattern (PAMP) molecules, and damage-associated molecular pattern molecules (DAMP), all of which aid in activating both innate and adaptive immune response [4].

4. Types of Naturally Derived Biomaterials Used in Cancer Treatment and Research

For cancer research and treatment, it is often more cost-effective, efficient, and safe to use biomaterials naturally synthesized by algae, plants, animals, and in some cases even bacteria, as opposed to synthetic materials [8]. Some of the more commonly utilized biomaterials include polysaccharides such as chitosan, noted for its cytotoxic ability in certain breast cancer lines; hyaluronic acid, which is significantly present in tumors and can function as an effective drug/adjuvant carrier; alginate, which induces the release of proinflammatory cytokines from macrophages; or pectin, which can function as a drug carrier and has apoptotic-stimulating abilities [2], [8].

Polar lipids can also be used for cancer treatment, as their amphipathic structure allows them to carry various drugs that are either hydrophobic or hydrophilic [2]. Lipid-based nanomaterials can serve as carriers of antigens and immune adjuvants targeting DCs, to initiate a CTL response [3]. The use of lipid-based biomaterials and liposomes showed to have a significant effect on the immune system. One study sought to create a cancer vaccine by loading liposomes with a peptide specific to melanoma, TRP2, and a CpG-ODN immune adjuvant, and the vaccine was found to improve the survival of tumor-bearing mice [15]. Some lipids even have direct anticancer properties, as in the case with alkyl phospholipids (APLs) which have shown antineoplastic activity by interfering with lipid metabolism in cancer cells [2].

Other natural biomaterials such as polypeptides and vitamin derivates can also be used as both carriers and tumor antagonists. Polypeptides and peptide derivatives can also serve as drug and gene therapeutics carriers and are recognized for their significant biocompatibility and chemical reactivity, and certain polypeptides such as polylysine, polyarginine, polyhistidine, and polyglutamic acid have direct antitumoral or immunomodulatory activities [2]. The vitamin E derivate alpha-tocopherol-succinate (TOS) can be paired with polysaccharide carriers such as hyaluronic acid and chitosan to form micelles for the loading of hydrophobic drugs [2], [16].

TABLE 1. Specific Functions of Biomaterials in Cancer Treatment

Biomaterial	Type	Carrier	Directly Anti-Tumoral	Immunomodulatory
Alginate	Polysaccharide	✓		✓
Alkyl phospholipid	Lipid	✓	✓	
α-TOS	Vitamin E derivate	✓	✓	
Chitosan	Polysaccharide	✓		✓
Gold	Metal nanoparticle	✓	✓	✓
Hyaluronic Acid	Polysaccharide	✓		
Iron Oxide	Metal nanoparticle	✓	✓	✓
Liposomes	Lipid	✓		✓
Pectin	Polysaccharide	✓	✓	
Polyarginine	Polypeptide	✓	✓	
Polyglutamic acid	Polypeptide	✓		✓
Polyhistidine	Polypeptide	✓	✓	
Polylysine	Polypeptide	✓	✓	
Silica	Nanoparticle	✓		✓

5. Curcumin Nanoparticles

One very notable example of a plant-extract biomaterial used to treat different cancers is curcumin, from which nanoparticles are created and then usually loaded onto a separate biomaterial carrier to then target tumors and cancer cells. Curcumin, chemically known as diferuloylmethane, is a polyphenolic substance extracted from the plant *Curcuma longa*, colloquially known as turmeric and used as a spice and coloring agent worldwide [17], [18], [19]. It has also been used in traditional Indian (Ayurvedic) and traditional Chinese medicine (TCM) practices to treat inflammation caused by a wide variety of disorders [19]. Research has shown that it suppresses NF-κB activation, a pathway that normally stimulates inflammation in cancer cells [18], [20], [21]. Curcumin also can downregulate AP-1 and STAT-3 pathways, which effectively retards the growth of cancer cells [18], [22].

Several studies have been performed to test if curcumin nanoparticles could truly work as anticancer agents when coupled with carriers formed from various biomaterials, and the results show promise for curcumin's use in the future for cancer treatment. Chitosan-coupled curcumin nanoparticles showcased the prominent mucoadhesive ability and drug retention when used to treat mouth cancer [23]. Another study used a silk fibroin and a blend of silk fibroin and chitosan polymers to encapsulate curcumin nanoparticles, where silk fibroin carriers were found to be more efficient due to their greater entrapment efficiency [24]. Other studies showed the prominent activity of curcumin nanoparticles against triple-negative breast cancer (TNBC) cells

and drug-resistant human ovarian adenocarcinoma cells [25]. Perhaps one of the most surprisingly effective uses of curcumin nanoparticles is against pancreatic cancer, and phase II *in vivo* studies reported that some patients experienced tumor regression and increased life expectancy, albeit small [20]. Numerous studies using curcumin against colorectal cancer reported that a variety of different carriers could be used to effectively transmit curcumin into the cell, including liposomes, other lipid carriers, micelles made of polymers, gold particles, and nanogels [22].

6. Conclusion

Various biomaterials could provide effective solutions to the massive challenges presented by cancer. By using polymeric scaffolds, tissue engineering, and intelligent biomaterial solutions such as organs-on-chips and microfluidics, tumor models can show a much more in-depth picture of the complexity of cancer dynamics and microenvironments. Many biomaterials themselves present antitumor or immunomodulating capabilities, which is a significant benefit when researchers look to use biomaterials therapeutically, in addition to being able to function as carriers for other anticancer therapies. Researchers have already looked into many different types of natural biomaterials, and one area of potential future research could be to examine how anticancer properties of different biomaterials might interact with one another to optimize cancer treatments. Another excellent focus area for future research would be to examine what kinds of cancers different biomaterials work best against. The wealth of research available showing the efficacy of biomaterials in both in vitro and in vivo treatments lends a lot of hope to a future where safe, effective, and even low-cost solutions for cancer exist, and people no longer need to fear cancer as a death sentence.

Appendix 1. List of abbreviations

CTL - cytotoxic T lymphocyte
DC - dendritic cell
Treg - T regulatory cell
MDSC - myeloid-derived suppressor cell
TME - tumor microenvironment
ECM - extracellular matrix
LOX - lysyl oxidase
PEI - polyethyleneimine
PEG - polyethylene glycol
APC - antigen-presenting cell
IFN-alpha – interferon-alpha
TLR - toll-like receptor
PAMP - pathogen-associated molecular pattern
DAMP - damage-associated molecular pattern
BMDC - bone marrow-derived dendritic cell
TNF - tumor necrosis factor
CpG-ODN - CpG oligodeoxynucleotide
alpha-TOS - alpha-tocopherol-succinate

7. References

[1] M. Vander Heiden and R. DeBerardinis, "Understanding the Intersections between Metabolism and Cancer Biology", Cell, vol. 168, no. 4, pp. 657-669, 2017. Available: https://www.sciencedirect.com/science/article/pii/S009286741631755X.
[2] Kinam Park, Biomaterials for cancer therapeutics, 2nd ed. Cambridge: Woodhead Publishing, 2020, pp. 1-80.
[3] Alhallak, K., Sun, J., Muz, B., & Azab, A. K. (2020). Biomaterials for cancer immunotherapy. In Biomaterials for Cancer Therapeutics (pp. 499–526). Elsevier. https://doi.org/10.1016/B978-0-08-102983-1.00018-1

[4] Bisht, S., Feldmann, G., Soni, S., Ravi, R., Karikar, C., Maitra, A., & Maitra, A. (2007). Polymeric nanoparticle-encapsulated curcumin ('nanocurcumin'): A novel strategy for human cancer therapy. Journal of Nanobiotechnology, 5(1), 3. https://doi.org/10.1186/1477-3155-5-3

[5] Bonnans, C., Chou, J., & Werb, Z. (2014). Remodelling the extracellular matrix in development and disease. Nature Reviews Molecular Cell Biology, 15(12), 786–801. https://doi.org/10.1038/nrm3904

[6] Dhillon, N., Aggarwal, B. B., Newman, R. A., Wolff, R. A., Kunnumakkara, A. B., Abbruzzese, J. L., Ng, C. S., Badmaev, V., & Kurzrock, R. (2008). Phase II Trial of Curcumin in Patients with Advanced Pancreatic Cancer. Clinical Cancer Research, 14(14), 4491–4499. https://doi.org/10.1158/1078-0432.CCR-08-0024

[7] Emami, J., Rezazadeh, M., Rostami, M., Hassanzadeh, F., Sadeghi, H., Mostafavi, A., Minaiyan, M., & Lavasanifar, A. (2015). Co-delivery of paclitaxel and α -tocopherol succinate by novel chitosan-based polymeric micelles for improving micellar stability and efficacious combination therapy. Drug Development and Industrial Pharmacy, 41(7), 1137–1147. https://doi.org/10.3109/03639045.2014.935390

[8] Goldberg, M. S. (2015). Immunoengineering: How Nanotechnology Can Enhance Cancer Immunotherapy. Cell, 161(2), 201–204. https://doi.org/10.1016/j.cell.2015.03.037

[9] Gupta, V. (2009). Fabrication and characterization of silk fibroin-derived curcumin nanoparticles for cancer therapy. International Journal of Nanomedicine, 115. https://doi.org/10.2147/IJN.S5581

[10] Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. Cell, 144(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

[11] Hao, S., Yan, Y., Ren, X., Xu, Y., Chen, L., & Zhang, H. (2015). Candesartan-graft-polyethyleneimine cationic micelles for effective co-delivery of drug and gene in anti-angiogenic lung cancer therapy. Biotechnology and Bioprocess Engineering, 20(3), 550–560. https://doi.org/10.1007/s12257-014-0858-y

[12] Jo, Y., Choi, N., Kim, K., Koo, H.-J., Choi, J., & Kim, H. N. (2018). Chemoresistance of Cancer Cells: Requirements of Tumor Microenvironment-mimicking In Vitro Models in Anti-Cancer Drug Development. Theranostics, 8(19), 5259–5275. https://doi.org/10.7150/thno.29098

[13] Kretlow, J. D., & Mikos, A. G. (2008). From material to tissue: Biomaterial development, scaffold fabrication, and tissue engineering. AIChE Journal, 54(12), 3048–3067. https://doi.org/10.1002/aic.11610

[14] Lai, C., Duan, S., Ye, F., Hou, X., Li, X., Zhao, J., Yu, X., Hu, Z., Tang, Z., Mo, F., Yang, X., & Lu, X. (2018). The enhanced antitumor-specific immune response with mannose- and CpG-ODN-coated liposomes delivering TRP2 peptide. Theranostics, 8(6), 1723–1739. https://doi.org/10.7150/thno.22056

[15] Mazzarino, L., Loch-Neckel, G., Bubniak, L. D. S., Mazzucco, S., Santos-Silva, M. C., Borsali, R., & Lemos-Senna, E. (2015). Curcumin-Loaded Chitosan-Coated Nanoparticles as a New Approach for the Local Treatment of Oral Cavity Cancer. Journal of Nanoscience and Nanotechnology, 15(1), 781–791. https://doi.org/10.1166/jnn.2015.9189

[16] Park, O., Yu, G., Jung, H., & Mok, H. (2017). Recent studies on micro-/nano-sized biomaterials for cancer immunotherapy. Journal of Pharmaceutical Investigation, 47(1), 11–18. https://doi.org/10.1007/s40005-016-0288-2

[17] Pradhan, S., Hassani, I., Clary, J. M., & Lipke, E. A. (2016). Polymeric Biomaterials for In Vitro Cancer Tissue Engineering and Drug Testing Applications. Tissue Engineering Part B: Reviews, 22(6), 470–484. https://doi.org/10.1089/ten.teb.2015.0567
[18] Singh, A., Brito, I., & Lammerding, J. (2018). Beyond Tissue Stiffness and Bioadhesivity: Advanced Biomaterials to Model Tumor Microenvironments and Drug Resistance. Trends in Cancer, 4(4), 281–291. https://doi.org/10.1016/j.trecan.2018.01.008
[19] Vander Heiden, M. G., & DeBerardinis, R. J. (2017). Understanding the Intersections between Metabolism and Cancer Biology. Cell, 168(4), 657–669. https://doi.org/10.1016/j.cell.2016.12.039
[20] Wang, C., Ye, Y., Hu, Q., Bellotti, A., & Gu, Z. (2017). Tailoring Biomaterials for Cancer Immunotherapy: Emerging Trends and Future Outlook. Advanced Materials, 29(29), 1606036. https://doi.org/10.1002/adma.201606036
[21] Wong, K. E., Ngai, S. C., Chan, K.-G., Lee, L.-H., Goh, B.-H., & Chuah, L.-H. (2019). Curcumin Nanoformulations for Colorectal Cancer: A Review. Frontiers in Pharmacology, 10, 152. https://doi.org/10.3389/fphar.2019.00152
[22] Xia, Y., Shen, S., & Verma, I. M. (2014). NF-κB, an Active Player in Human Cancers. Cancer Immunology Research, 2(9), 823–830. https://doi.org/10.1158/2326-6066.CIR-14-0112
[23] Yang, F., Shi, K., Jia, Y., Hao, Y., Peng, J., & Qian, Z. (2020). Advanced biomaterials for cancer immunotherapy. Acta Pharmacologica Sinica, 41(7), 911–927. https://doi.org/10.1038/s41401-020-0372-z
[24] D. Hutmacher, "Biomaterials offer cancer research the third dimension", Nature Materials, vol. 9, no. 2, pp. 90-93, 2010. Available: http://www.australianprostatecentre.org/files/dietmar-hutmacher-article_2.
[25] N. Vallianou, A. Evangelopoulos, N. Schizas and C. Kazazis, "Potential Anticancer Properties and Mechanisms of Action of Curcumin", Anticancer research, vol. 35, no. 2, pp. 645-651, 2021. Available: https://ar.iiarjournals.org/content/anticanres/35/2/645.full.pdf. [Accessed 28 May 2021].
[26] A. Chevallier, Encyclopedia of herbal medicine, 3rd ed. New York: Dorling Kindersley Ltd, 2016, pp. 1-98. Available: https://toniau.ac.ir/med/wp-content/uploads/docs/Book/%DA%A9%D8%AA%D8%A7%D8%A8/Encyclopedia%20Herbal%20Medicine.pdf
[27] Maya & Rajsekhar, Sharadha & Rajsekhar, Vardhini. (2015). Curcumin Nanoparticles: A Therapeutic Review. Research Journal of Pharmaceutical, Biological and Chemical Sciences. 6. 2015-6.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

The author would like to acknowledge Assist. Prof. Elnur Tahirovic and Adna Sijercic, MSc, for their advice and input on this paper.

Analysis of High School Graduate Data Using Database Analytics Tools

Ezana Ćeman*, Ajdin Salihović*, Samed Jukić*
*International Burch University,
Sarajevo, Bosnia and Herzegovina
ezana.ceman@stu.ibu.edu.ba
ajdin.salihovic@stu.ibu.edu.ba
samed.jukic@ibu.edu.ba

Original research

Abstract: *It can be confidently stated that access to education is one of the most prized possessions available to us today. Although there are underlying factors such as the discrepancies in the education being provided worldwide, it is imperative that data scientists and all those interested take advantage of the data publicly available to draw necessary insights into how to better the education sector in our respective countries. The purpose of this research is to showcase various analytical insights into the 2020 New York State (NYS) high school graduation rate data using various advanced database systems techniques, specifically using SQL. With these analyses, further studies and conclusions can be drawn for local governments to implement into their plans to increase the quality of the schooling system, to aim for equality for all without regard to cultural and ethnic background, and to find discrepancies within the current system.*

Keywords: **Database, data analysis, graduate, high school, New York State, SQL.**

1. Introduction

The realm of education is one topic always of discussion due to the worldwide discrepancies in the quality of education being provided. Although education is every human being's right, access to basic education is a duty that must be fulfilled within countries worldwide. “Governments are typically expected to ensure access to basic education, while citizens are often required by law to attain education up to a certain basic level [1].” Although every country has its education system, what the world lacks is one uniform system, ensuring equal access to education for all. What Figure 1 below shows us is government expenditure per student as a percentage of GDP per capita, in terms of secondary education by country [2]. We can see that the top four countries with the highest expenditure rates are: The United Kingdom, Japan, the United States, and India.

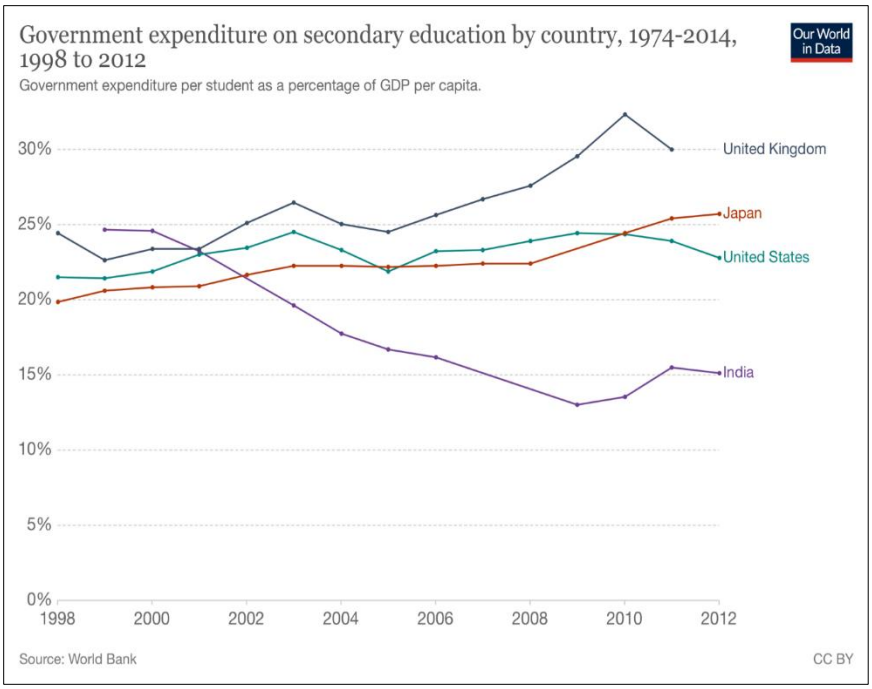


FIGURE 1. Government expenditure on secondary education by country, 1974-2014, 1998 to 2012 - *Source: Our World in Data*

Aside from government expenditures into the secondary education sector, gross enrollment ratios are imperative to analyze as enrollment of young men and women in secondary education, regardless of age statistics, can provide us with better insights into which countries are currently ranked with the highest values. Figure 2 below shows the world's leaders in gross enrollment in secondary education ranging from 1970 to 2014 with Portugal and Barbados ranking the highest with heavily indebted poor countries (HIPC) ranking the lowest [2].

We will admit that these results did come as a shock to us as the first instinct is almost always for our minds to look to world superpowers for the highest enrollment ratios. As we can see, this predisposition is false when looking at the data.

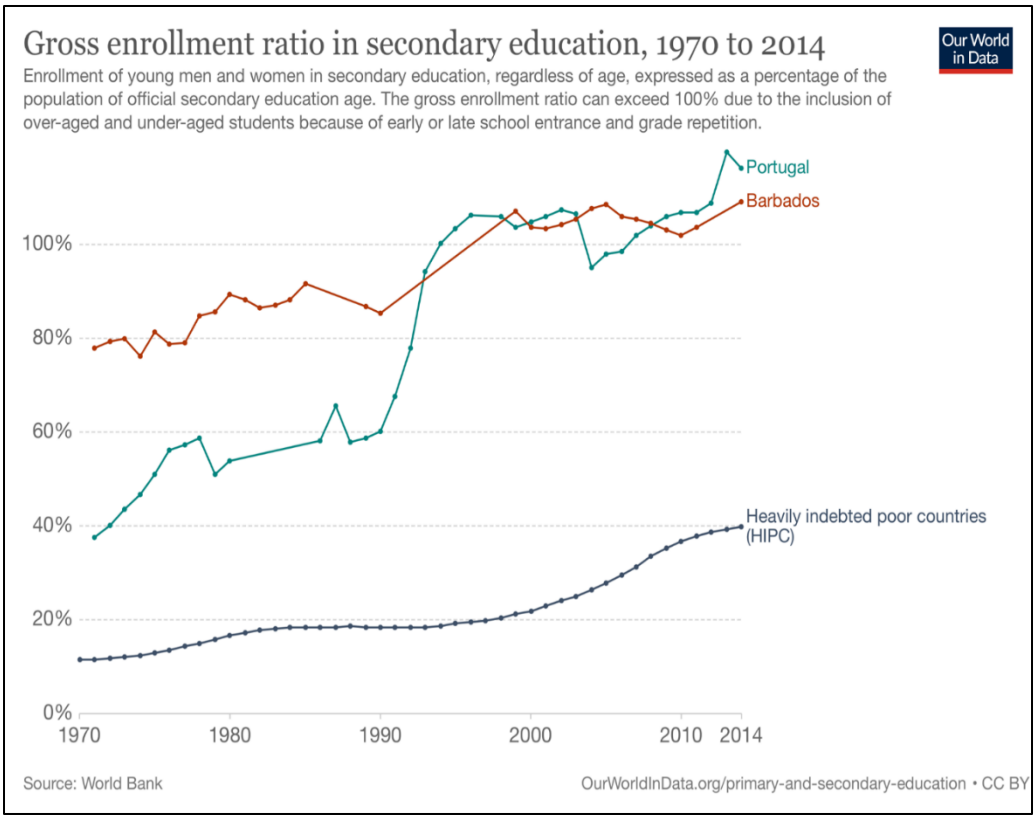


FIGURE 2. Gross enrollment ratio in secondary education, 1970 to 2014
- *Source: Our World in Data*

For this research, the region of the world that will be focused on pertains to the northeast United States, specifically the state of New York. NYS currently has 731 districts, 4,421 public schools, and 351 charter schools actively functioning at the moment [3]. Out of all the 4,421 public schools, there are a total of 2,053 high schools in NYS, including 1,520 public schools and 533 private schools [4]. As of June 30th, 2019, there were a total of 2,598,921 K-12 public school students in NYS [3]. From 2019-20, 815,707 students were recorded as high school students, ranging from grades 9, 10, 11, 12, and ungraded secondary which is the standard secondary school grades in NYS [5]. An additional grade level, denoted 'Ungraded Secondary' represents specialized high schools in NYS that do not follow the standard grading system. However, these students fall under the category of high school students nonetheless.

As ethnicity will play a significant role in this research, it is worth noting the latest available data regarding enrollment by ethnicity. Figure 3 below represents the 2019-20 enrollment count based on ethnicity [5]. White-identifying students made up the highest group, with an estimated 353,000 high school students. Hispanic or Latino high school students accounted for an estimated 219,000 students. Black or African American high school students totaled 141,000. Asian or Native Hawaiian/Other Pacific Islander high school students totaled 81,000. Multiracial high school students accounted for a total of 17,000 and lastly, there were a recorded 6,000 American Indian or Alaska Native high school students.

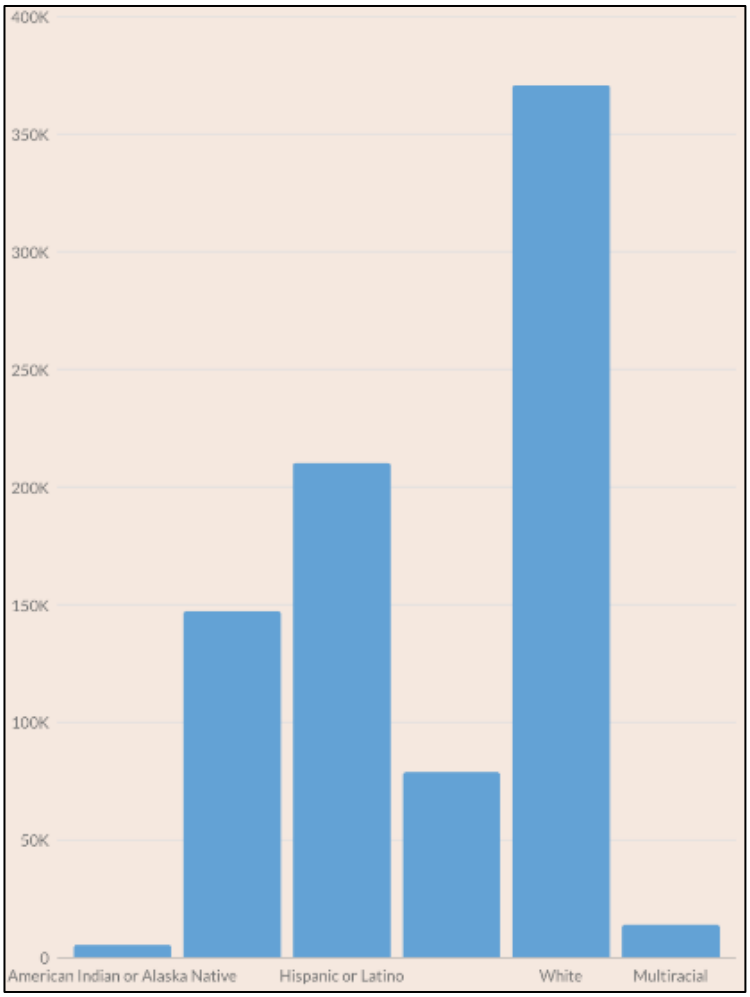


FIGURE 3. Enrollment by Ethnicity, 2019-20 - *Source: The New York State Education Department*

In addition, it is worth noting the latest data available in terms of enrollment by grade to provide us with an overview of how many students are in each of the high school grades relating to this study. Table 1 below represents the data available from 2019-20 in terms of enrollment by grade [5].

Above we can see that there is an about-even split between grades 9 – 12 regarding enrollment by grade.

There is a multitude of database analytics tools available for use with the main purpose of utilization being for data analysis. Not every tool fits every single need however, great insights can be found through researching which tool(s) work the best for the respective field of research. The field of data analytics has been growing day by day as new insights are being pulled from data in endless ways imaginable. The purpose of this research is to showcase various data analyses through the use of database analytics tools by utilizing comparative study techniques in terms of graduation percentages. After finding the right dataset to work with, various steps were taken in the preprocessing of the dataset to ensure that all of the data is being analyzed as accurately as possible. A variety of exploratory and description data analyses were conducted on the 2020 NYS graduation rates dataset to find various comparisons that can be implemented further in additional research.

TABLE 1. Enrollment by Grade, 2019-20
Source: The New York State Education Department

9th Grade	
211,978	26%
10th Grade	
203,562	25%
11th Grade	
191,168	23%
12th Grade	
189,493	23%
Ungraded Secondary	
19,506	2%

2. Methods and Materials

Dataset Overview

The New York State Education Department (NYSED) frequently publishes data in regards to a multitude of aspects of the education sector in the state for public use, which is where we found the dataset to be used in this research. Titled "GRAD_RATE_AND_OUTCOMES_2020", this dataset features a total of 227,451 rows and a total of 37 columns [6]. This gives us an estimated 8,415,687 numerical and categorical values to work with. This dataset showcases "outcomes of designated subgroups are reported by the total public school (aggregated data for all districts and charter schools), county (aggregated data for all districts and charter schools in the county), Needs-to-Resource-Capacity (N/RC) group, district, and public schools [6]."

One of the most significant indicators that will be the main priority of this research pertains to 'SUBGROUP_NAME', which consists of a text value and 25 differentiating characteristics. NYS has a standard definition of the subgroups used within their documentation which is shown in Table 2 below.

TABLE 2. Subgroups (NYS)

1. 'All Students'	13. 'English Language Learner'
2. 'Female'	14. 'Formerly English Language Learner'
3. 'Male'	15. 'Economically Disadvantaged'
4. 'American Indian/Alaska Native'	16. 'Not Economically Disadvantaged'
5. 'Black'	17. 'Migrant'
6. 'Hispanic'	18. 'Not Migrant'
7. 'Asian/Pacific Islander'	19. 'Homeless'
8. 'White'	20. 'Not Homeless'
9. 'Multiracial'	21. 'In Foster Care'
10. 'General Education Students'	22. 'Not in Foster Care'
11. 'Students with Disabilities'	23. 'Parent in Armed Forces'
12. 'Not English Language Learner'	24. 'Parent Not in Armed Forces'.

Each subgroup has its corresponding 'SUBGROUP_CODE', consisting of a specific 2-digit code identifying the demographic at hand.

Utilized Technologies

For the purpose of this research, due to the large size of the dataset being analyzed, we found that through the use of the programming language SQL, we would be able to acquire effective results while using various advanced database system techniques. SQL stands for Structured Query Language and it is "a database computer language designed for the retrieval and management of data in a relational database [7]." We have found through trial and error by testing out other database analytics tools, that we can acquire the greatest results by using the SQL language. By creating our tailored queries, we were able to retrieve various insights into the 2020 graduate rate dataset which in turn will lead us to our conclusions which are noted towards the end of this work. In addition, the Microsoft Excel program was utilized to create the visuals relating to the results of this research.

Data Cleaning and Preprocessing

For this research, cleaning is the process of detection and correcting records from records set, table, or database. For example, if there are incomplete, incorrect, inaccurate, or irrelevant, we would modify or delete the faulty data. In the case of discrepancies in the education being provided worldwide, in this case for the New York State (NYS) high school graduation rate. We might see a discrepancy between students enrolled versus students who graduated in case of seeing more graduates than the enrolled students of that class. We would check the number of students that repeated the year with the number of enrolled students for an example of class 2020. If a certain number of students of class 2019 have repeated the year they would also be graduating at the same time year after. This filtering is used when using WHERE when using SQL in which we can filter records and add more using AND until we are ready to run the query.

3. Results

As with every dataset, significant conclusions can be drawn through the use of statistical analysis techniques which was our first step in the data analysis process. Using both the 'SUBGROUP_CODE' and 'SUBGROUP_NAME' indicators as a basis for our queries, various statistical values were found which in turn helped bring light to the current demographic relationship to graduate rates. Based on these indicators, what was found was: the sum of the total enrolled based on the subgroup, the sum of total graduates based on the subgroup, the sum of total locals based on the subgroup, the sum of total registered based on the subgroup, the sum of total advanced registered based on the subgroup, the sum of total diploma credentials based on the subgroup, the sum of the total still enrolled based on the subgroup, the sum of total dropouts based on the subgroup, the average of enrolled based on the subgroup, the average of graduates based on the subgroup, the average of locals based on the subgroup, the average of registered based on the subgroup, the average of advanced registered based on the subgroup, the average of diploma credentials based on the subgroup, the average of still enrolled based on the subgroup, and the average enrolled based on the subgroup.

From these analyses, the most relevant results found are summarized below.

Figure 4 shows that students in the Non-Migrant, Not in Foster Care, Parent Not in Armed Forces, Not Homeless, and Not English Language Learners categories are the highest in terms of enrolled students.

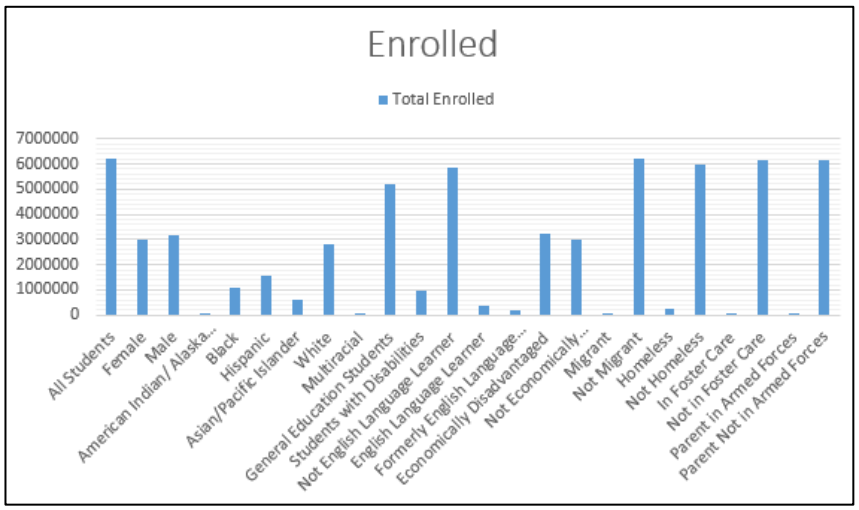


FIGURE 4. Calculates Sum of total Enrolled based on Subgroup.

Figure 5 shows that students in the Non-Migrant, Not in Foster Care, Parent Not in Armed Forces, Not Homeless, and Not English Language Learners categories are the highest in terms of graduated students.

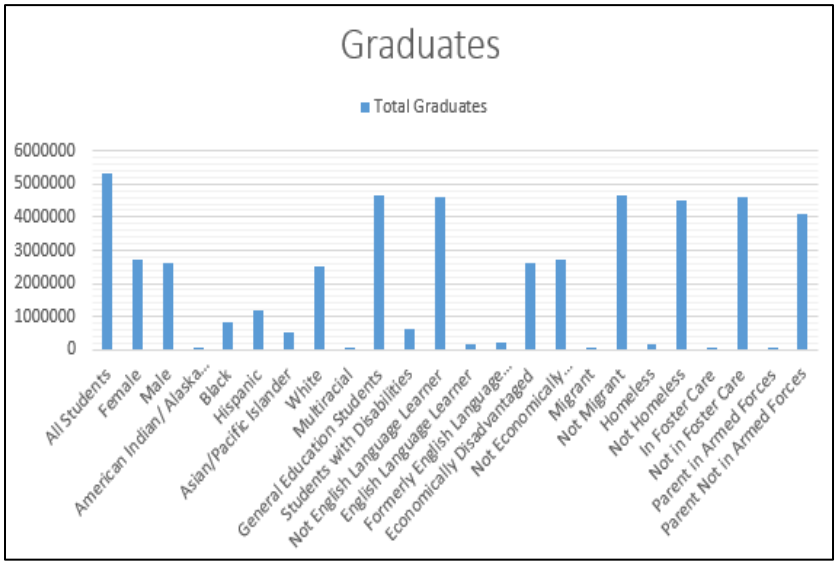


FIGURE 5. Calculates Sum of total Graduates based on Subgroup.

Figure 6 shows us that students in the Non-Migrant, Not in Foster Care, Parent Not in Armed Forces, Not Homeless, and Not English Language Learners categories are the highest in terms of local students. As can be seen, there is a recurring trend here thus far.

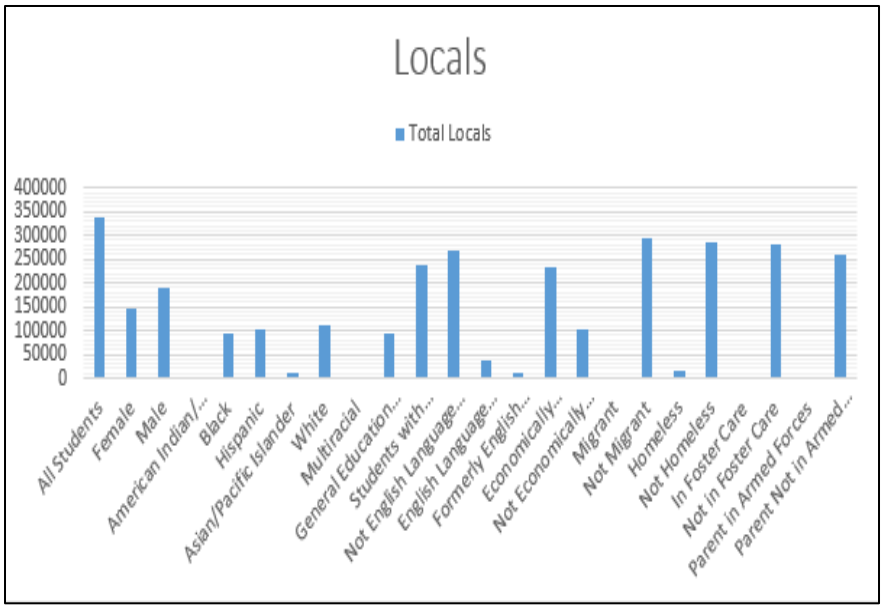


FIGURE 6. Calculates Sum of total Locals based on Subgroup.

Figure 7 shows us a representation of the sum of total registered students based on their respective subgroups. Similar to Figures 4 – 6, there is a recurring theme present.

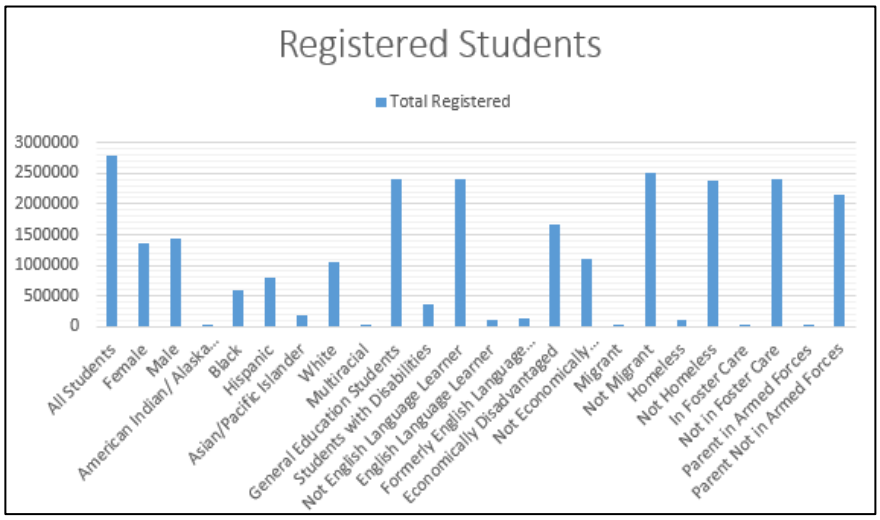


FIGURE 7. Calculates Sum of total Registered based on Subgroup.

Figure 8 shows us that General Education Students followed by Not Migrant, Not Homeless, and Not in Foster Care make up the highest groups for advanced registered students. In NYS, advanced registered stands for students who are taking college-level classes at a high school level.

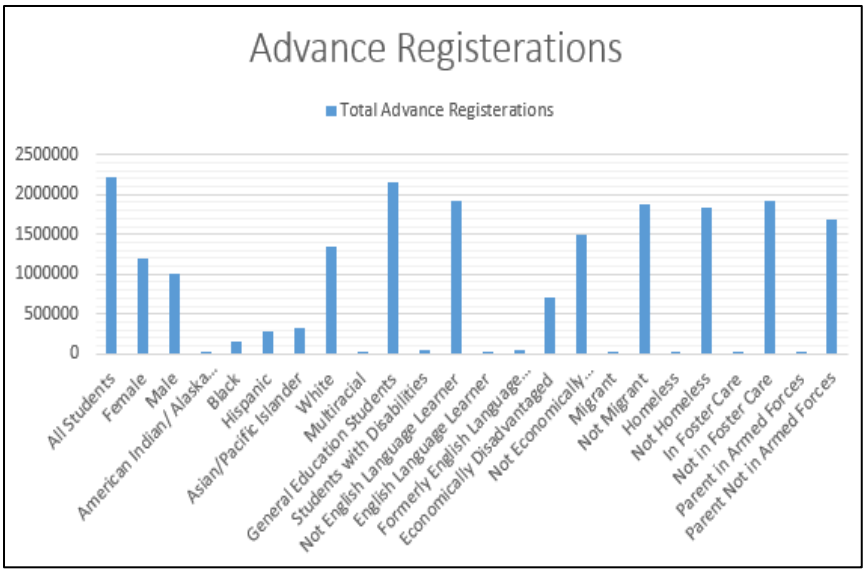


FIGURE 8. Calculates Sum of total Advance Registered based on Subgroup.

Figure 9 shows us that Students with Disabilities followed by Not Migrant, Not Homeless, and Not in Foster Care make up the highest groups for total diploma credentials.

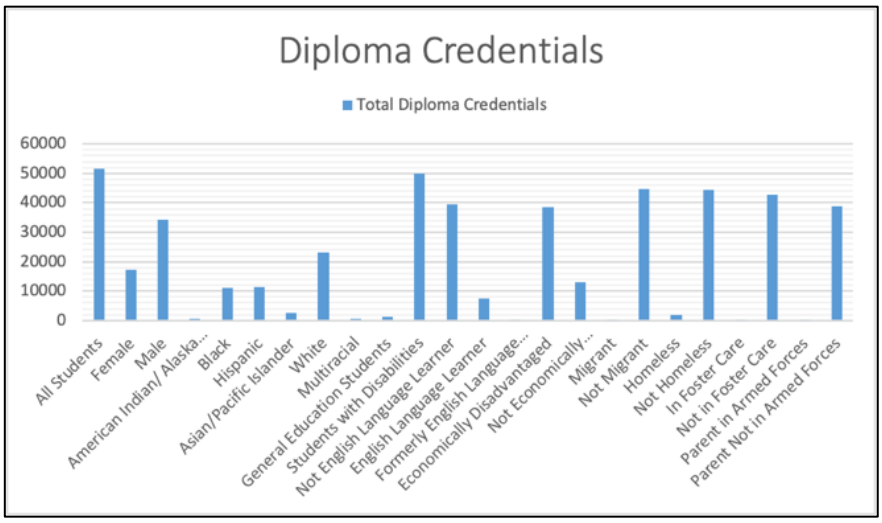


FIGURE 9. Calculates Sum of total Diploma Credentials based on Subgroup

Figure 10 shows us a graphical representation of the sum of the total still enrolled students based on the subgroup. A recurring theme to the previous figures remains.

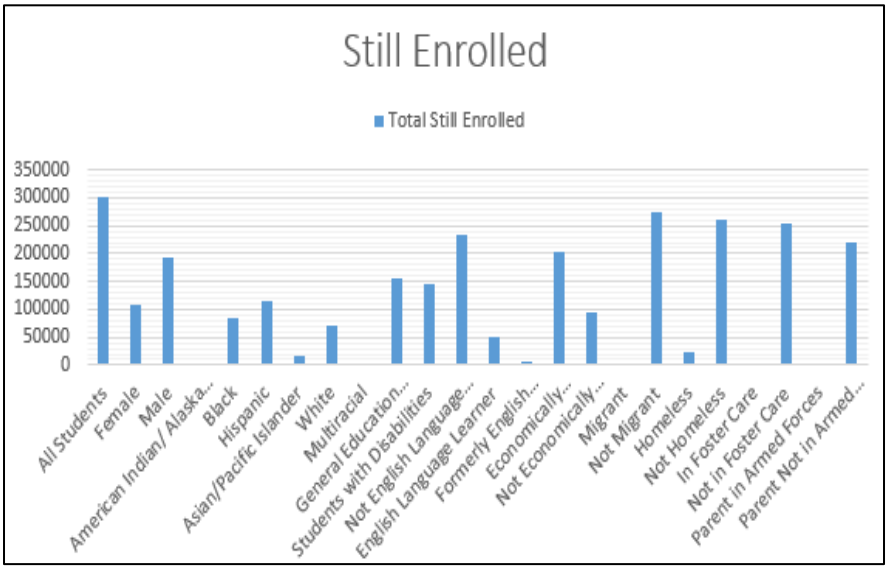


FIGURE 10. Calculates Sum of total Still Enrolled based on Subgroup.

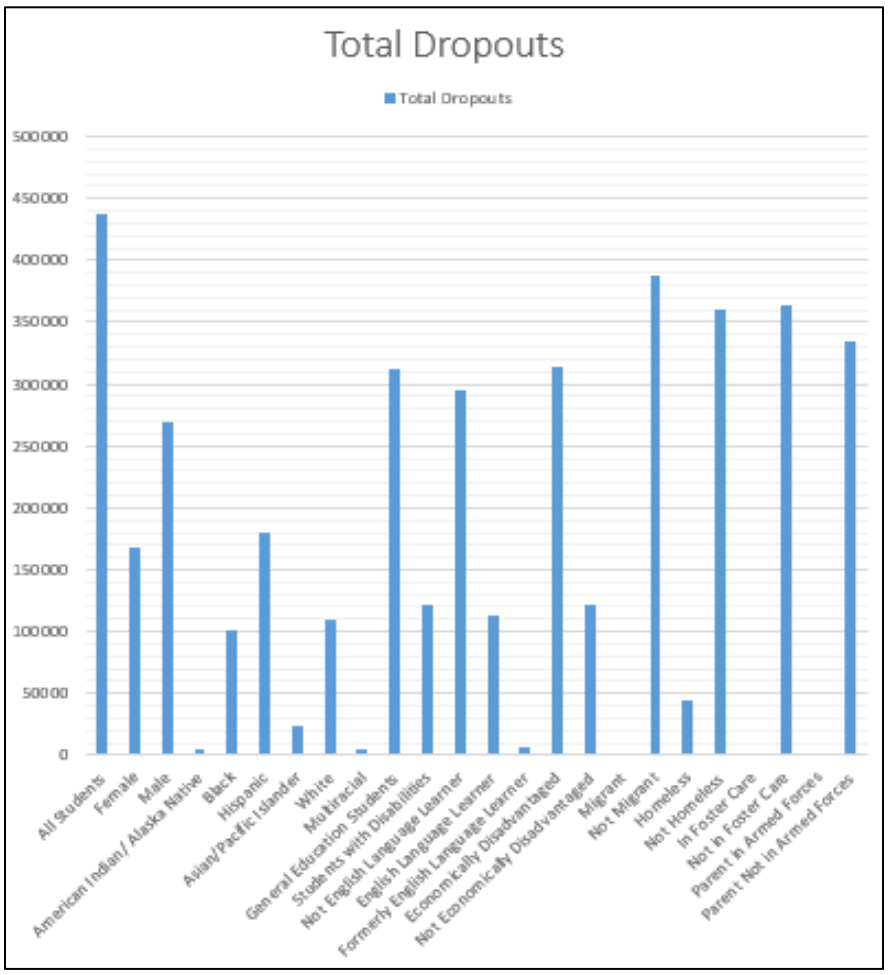


FIGURE 11. Calculates Sum of total Dropouts based on Subgroup.

Following these relevant sums and averages computed, we wanted to showcase the statistics of total dropped out students county-wise for those counties which have more than a total of 10,000 dropout students. For reference, NYS is broken up into 62 different counties which are all included within the dataset as well.
Figure 12 below was created which represents the desired output.

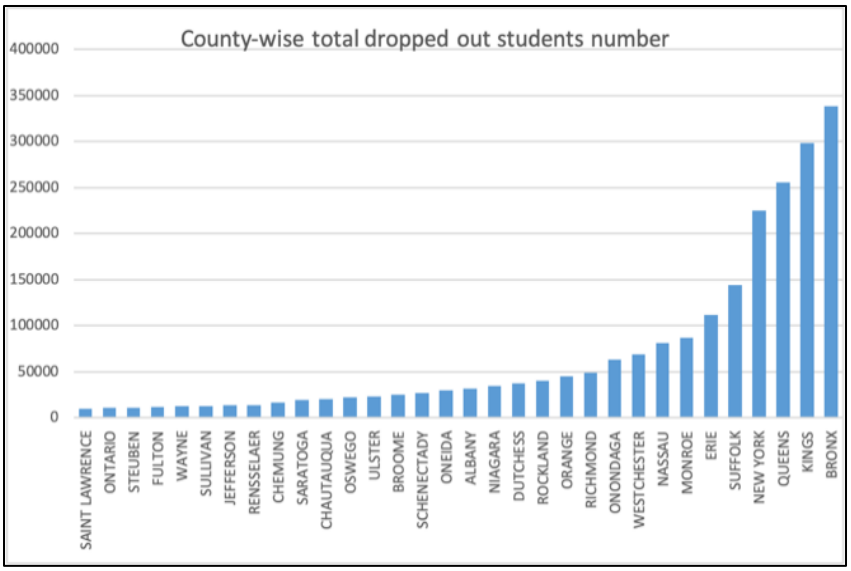


FIGURE 12. The county-wise total dropped out students.

The graph clearly shows that Bronx county has the maximum number of dropout students whereas Saint Lawrence county has the minimum among the counties that have more than 10,000 dropout students.
Following the computation of county-wise data, what follows is the statistics of graduation percentages statewide over the various subgroups for the '2014 Total Cohort - 6 Year Outcome' membership.
Figure 13 below was created which represents the desired output.

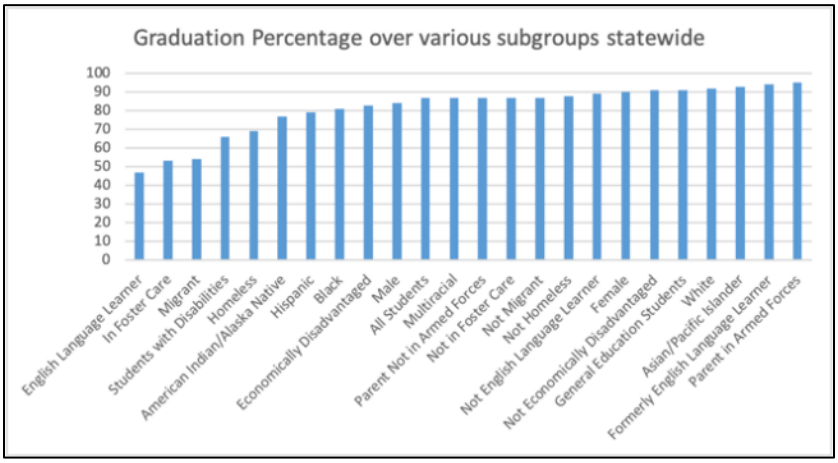


FIGURE 13. Graduation percentages over various subgroups.

By analyzing the output data, it was found that English Language Learners have the lowest percentage of graduation whereas students whose parents are in the Armed Forces have

the highest percentage of graduation. It's tough for a nonnative English Language Learner to get used to an entirely new language and way of life so it may be a reason for the lowest percentage. The students whose parents are in the Armed Forces may have better government-provided facilities which is a catalyst for their highest percentage. The graph also shows that female students have a higher percentage of graduating compared to male students. English Language Learners, In Foster Care, and Migrant subgroups students should be given special care because their graduation percentage is the lowest and below 60. Following the aforementioned analysis, we wanted to list the counties and number of schools in each of them where no students graduated in various subgroups over the cohort '2014 Total Cohort - 6 Year Outcome'. Only those county schools are to be considered where the Need to Resource Capacity category is 'Urban-Suburban High N/RC Districts'.

The results have been visualized in Figure 14 below.

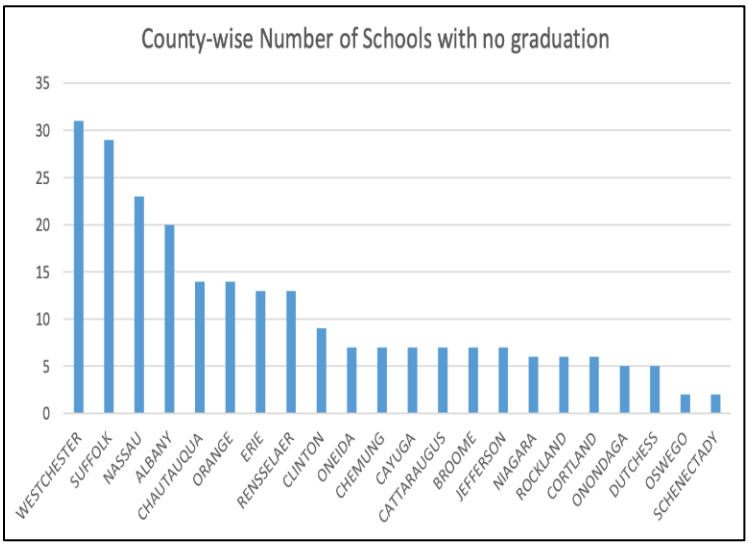


FIGURE 14. County-wise Number of Schools with no graduation.

By analyzing the above data, we see that Westchester County has the highest number of schools where no students in various categories graduated in the last six years.

Next, a comparative study was created between the graduation percentage of all students vs. black students over various counties where at least 100 black students were enrolled during the '2014 Total Cohort - 6 Year Outcome' period.

The results have been visualized in Figure 15 below.

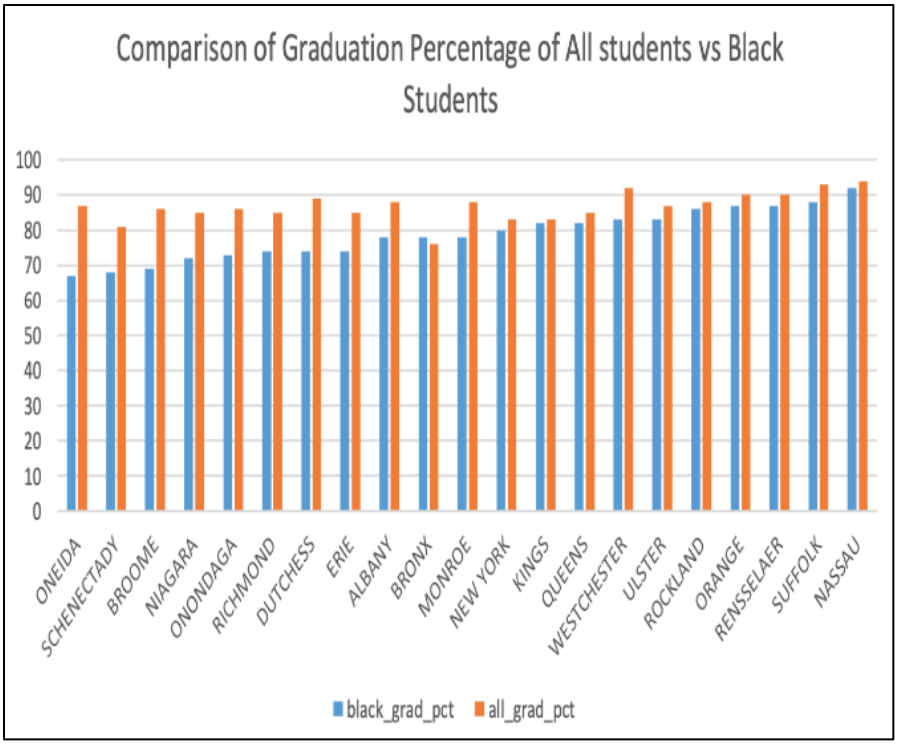


FIGURE 15. Comparison of Graduation Percentages of All Students vs. Black Students.

The graph shows that almost every county graduation percentage for Black students is smaller than the percentage of All Students. Only Bronx County is different where the percentage is greater for Black students than that of All Students. The demographic breakdown for Bronx County is 43.6% Black or African American [8].

Lastly, the final analysis conducted in this study was to find the dropout percent of students based on the 'Need to Resource Capacity' category for the 'All Students' subgroup and the cohort '2016 Total Cohort - 4 Year Outcome - August 2020'.

The tabular output is depicted in Table 3 below.

TABLE 3. Dropout percentages based on selected indicators.

aggregation_name	dropout_pct
NRC: Low Needs	1%
NRC: Charters	3%
NRC: Average Needs	4%
NRC: NYC	6%
NRC: Rural High Needs	7%
NRC: Buffalo, Rochester, Yonkers, Syracuse	10%
NRC: Urban-Suburban High Needs	11%

The tabular output shows that the NRC: Low Needs category has the minimum dropout percentage and the NRC: Urban- Suburban High Needs category has the maximum.

4. Discussion and Conclusions

Throughout this research journey, various advanced database systems techniques were utilized to be able to provide insights into the graduation rate dataset provided to the public by the New York State Department of Education. The application of various data analytic processes was successfully done here through various steps along the way: data collection, data cleaning, data preprocessing, data analysis, and both data and results interpretation. There were a few minor obstacles that came in the way when it came to our implementation of the aforementioned steps yet results were found, analyzed, and interpreted.

Here is a summary of the results gathered from this research study:

1. Students in the Non-Migrant, Not in Foster Care, Parent Not in Armed Forces, Not Homeless, and Not English Language Learners categories are the highest in terms of enrolled students.
2. Students in the Non-Migrant, Not in Foster Care, Parent Not in Armed Forces, Not Homeless, and Not English Language Learners categories are the highest in terms of graduated students.
3. Bronx county has the maximum number of dropout students whereas Saint Lawrence county has the minimum among the counties that have more than 10,000 dropout students.
4. English Language Learners have the lowest percentage of graduation whereas students whose parents are in the Armed Forces have the highest percentage of graduation.
5. Westchester County has the highest number of schools where no students in various categories graduated in the last six years.
6. Almost every county graduation percentage for Black students is smaller than the percentage of All Students. Only Bronx County is different where the percentage is greater for Black students than that of All Students (due to current demographics).

This topic was an interesting one that has peaked our internet for potential further research opportunities. There are various ways to go about interpreting data for usable results and this research paper just shows a few methods using the tools available to us to do so. The possibilities are endless with data science, and hopefully, through the use of this work, conclusions will be drawn to ensure differences are made within local, state-wide, and national governments to push for equality for all in terms of accessibility and quality of information.

7. References

[1] Roser, M. (2016, August 31). Global Education. Our World in Data. https://ourworldindata.org/global-education

[2] Roser, M. (2013, July 17). *Primary and Secondary Education*. Our World in Data. https://ourworldindata.org/primary-and-secondary-education

[3] NYSED Data Site. (n.d.). https://data.nysed.gov

[4] High Schools (h.s.). New York High Schools. https://high-schools.com/directory/ny/

[5] 2018 | NY STATE - Enrollment Data | NYSED Data Site. (n.d.). NYSED Data Site. Retrieved June 1, 2021, from https://data.nysed.gov/enrollment.php?state=yes&year=2018&grades%5B%5D=09&grades%5B%5D=10&grades%5B%5D=11&grades%5B%5D=12&grades%5B%5D=14

[6] https://data.nysed.gov/files/gradrate/19-20/gradrate.zip

[7] SQL Tutorial - Tutorialspoint. (n.d.). Tutorialspoint. Retrieved June 1, 2021, from https://www.tutorialspoint.com/sql/index.htm

[8] Bronx County, NY. (n.d.). Data USA. Retrieved June 1, 2021, from https://datausa.io/profile/geo/bronx-county-ny

[9] National Research Council and National Academy of Education. (2011). High School Dropout, Graduation, and Completion Rates: Better Data, Better Measures, Better Decisions. Committee for Improved Measurement of High School Dropout and Completion Rates: Experuidance on Next Steps for Research and Policy Workshop. R.M. Hauser and J.A. Koenig, Editors. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies PrPress.

[10] Anderson, K.P., and Ritter, G.W. (2017). Disparate use of exclusionary discipline: Evidence on inequities in school discipline from a U.S. state. Education Policy Analysis Archives, 25(49).

[11] Balfanz, R., Herzog, L., and MacIver, D.J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. Educational Psychologist, 42(4), 223-235.

[12] American Educational Research Association. (2015). AERA statement on the use of value-added models (VAM) for the evaluation of educators and educator preparation programs. Educational Researcher 44(8), pp. 448-452. Available: https://journals.sagepub.com/doi/10.3102/0013189X15618385 [August 2019].

[13] Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. Educational Researcher, 36(5), 258-267.

[14] National Academies of Sciences, Engineering, and Medicine. 2019. Monitoring Educational Equity. Washington, DC: The National Academies Press. doi: 10.17226/25389.×

[15] Aucejo, E., and Romano, T.F. (2016). Assessing the effect of school days and absences on test score performance. Economics of Education Review, 55, 70-87. Available: https://doi.org/10.1016/j.econedurev.2016.08.007 [March 2019]

Frequency Locked Loop in Estimating Active, Reactive, and Apparent Powers

Meris Lihić[1], Slobodan Lubura[2]
[1]International Burch University, Sarajevo, Bosnia and Herzegovina
[2]University of East Sarajevo, East Sarajevo, Bosnia and Herzegovina
meris.lihic@stu.ibu.edu.ba
slobodan.lubura@ibu.edu.ba

Original research

Abstract: *In this paper, a new power calculation method has been presented. This method is based on a second-order generalized integrator frequency locked loop (SOGI-FLL) and has enhanced features over classical methods for power calculation widely used in industry. The FLLs have a wide variety of applications such as power converters, grid synchronization, sensorless flux estimation, and control of motor drives. The nature of the FLL allows for it to be a potentially perfect calculation method for power calculation. The obtained results showcase the effectiveness of the proposed power calculation method.*

Keywords: Power measurement, addition theorems for trigonometry, DC component elimination, gain normalization, grid frequency synchronization, MAF filters, MATLAB/Simulink.

1. Introduction

Renewable energies have become important due to the classic electricity infrastructure turning towards a distributed power generation system [1]. Thus, the role of an interface between generation systems and the electricity network will be filled in by the electricity networks of the future, which will extensively use power electronic devise, information and communication technology applications [2]. The power converters, grid-connected, must be carefully designed and controlled to achieve an optimal and efficient operation in distributed power generation systems [3].

One important issue in this distributed power generation system is power measurement in each internal connecting point of the power system. Typically, power calculation methods inherently have built-in zero-crossing detection (ZCD) of grid voltage and current. But, ZCD suffers from voltage (current) sags, drops, spikes and it is not a reliable scheme. As consequence, the power calculation method fails.

In this paper, a new power calculation method has been presented. This method is based on Frequency Locked Loop (FLL) and has enhanced features over classical methods (ZCD) for power calculation widely used in industry. The FLL is a non-linear closed-loop system that can be used in a wide variety of applications such as grid synchronization, flux estimation, and control of motor drives without using sensors [4].

2. Methodology

The first stage of this paper was based on state of art. Due to the nature of the paper, proper research had to be performed, as the FLLs are a rather unexplored topic. Such issues are the complexity of the proposed algorithms, DC offset which makes approximations impossible, the sensitivity of integral-based calculations to external sources, which are the focus of this paper.

The second stage was to develop a mathematical model for the issued problem. The power calculation algorithm is based on addition theorems for trigonometry. By separating the voltage and current into respective *v's* and I've components by using the FLL, and applying addition theorems to them, a simple way to calculate active and reactive power is achieved.

The third stage was to perform simulations using the Matlab/Simulink software. Simulation results confirmed the effectiveness for both the power measuring algorithm, as well as FLL in the task of estimating rapid frequency change.

The fourth stage is to prepare the system for implementation onto an FPGA board. This step has yet to be performed, due to its nature and size, and is intended to be performed in future work.

3. Proposed Power Calculation Method Based on SOGI-FLL

SOGI-FLL Short Overview

The generalized integrator (GI) is the basis in the implementation of most FLLs. Roughly speaking, the GI's structure is based on a double integrator which provides an infinite gain at its resonant frequency and behaves as the amplitude integrator of sinusoidal signals. There are various realizations of the GI. The most popular way of implementing the FLL-based synchronization techniques, which is the focus of this paper, is the second-order generalized integrator (SOGI) [5].

The SOGI-FLL is a simple, and yet valuable tool because, in addition to providing filtered in-phase and quadrature-phase versions of its input, it can directly estimate frequency, and indirectly the phase angle and amplitude of the signal. The SOGI-FLL, nevertheless, has a limited filtering capability. In other words, in the presence of DC offset, harmonics, and inter-harmonics,

the estimated quantities by the SOGI-FLL suffer from ripples [5].

The SOGI-FLL estimates the frequency of the input signal. The frequency of the input signal does not experience sudden changes. Consequently, the FLL algorithm shows greater performance, when the phase angle of the input signal changes, than its PLL-based counterpart [1]. This GI is based on the principle that the time-domain convolution product of a sinusoidal function, by itself, gives rise to the original function which is multiplied by the time variable. Therefore, a processing block with a transfer function that is equal to the Laplace transform of a sinusoidal function (i.e., a resonator), will behave as an "amplitude integrator" for a sinusoidal signal applied at the input. Additionally, the in-quadrature combination of the sine and cosine transfer functions leads to an "ideal integrator" which is independent of the phase angle of the sinusoidal input signal [1].

Proportional-resonant controllers are based on the GI. Also, the GI has been applied to adaptive filtering applications and the PLL implement structure of this filter is shown in Fig. 1, where it is seen that the resonance frequency of the second-order generalized integrator (OGI) is an external parameter called ω'.

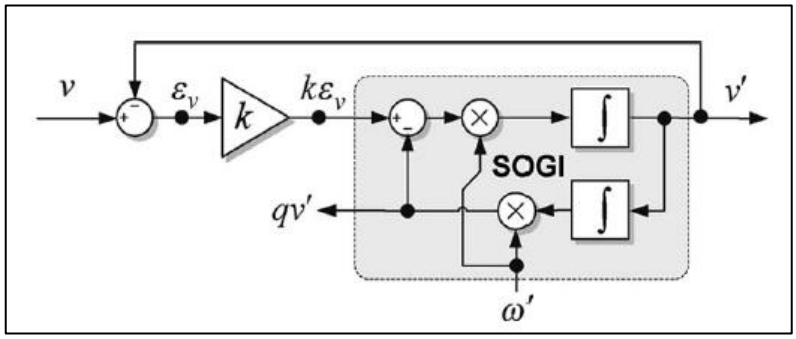


FIGURE 1. Block diagram of OGI [1].
The transfer function of the SOGI is given by:

$$\text{SOGI}(s) = \frac{v}{k\varepsilon_v}; (s) = \frac{\omega s}{s^2\omega'^2} \tag{1}$$

The resonance frequency is noted as ω', in general case, so it differs from the input frequency ω.

The two in-quadrature output signals of the adaptive filer in Fig. 1, i.e., v' and qv', are defined by the following transfer functions:

$$D(s) = \frac{v'}{v}(s) = \frac{k\omega' s}{s^2 + k\omega' s + \omega'^2} \tag{2a}$$

$$Q(s) = \frac{qv'}{v} = \frac{k\omega'^2}{s^2 + k\omega' s + \omega'^2} \tag{2b}$$

As (2a) shows, the bandwidth of the bandpass filter is determined by the gain k and is independent of the central frequency ω'. The same happens with the low-pass filter of (2b), in which the static gain only depends on gain k [1].

The figure below represents an exemplary SOGI-FLL synchronization system. The input grid frequency ω is directly detected by the FLL. On the other hand, the phase angle and amplitude of the input ought to be calculated indirectly.

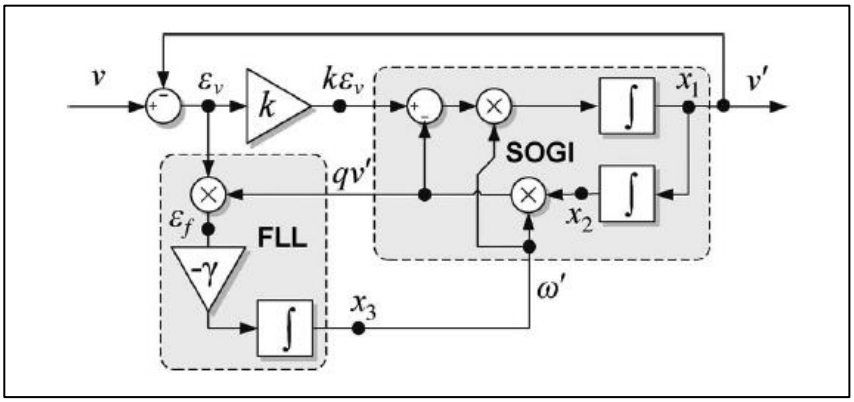


FIGURE 2. SOGI-FLL, a single-phase grid synchronization system [1].

The transfer function from the input signal v to the error ε_v is given by:

$$E(s) = \frac{\varepsilon_v}{v}(s) = \frac{s^2 + \omega'^2}{s^2 + k\omega' s + \omega'^2} \tag{3}$$

A frequency error ε_f is defined as the product of qv' by ε_v. The average value of the error ε_f will be positive when $\omega < \omega$', zero when $\omega = \omega$', and negative when $\omega > \omega$', where are ω' - resonance frequency (eg. 50 Hz) and ω – input grid frequency. As shown in Fig. 2, the DC component of the estimated frequency error can be made zero by shifting the SOGI resonance frequency ω' until it matches the grid frequency on the input ω. This is achieved by an integrated controller with a negative gain $-\gamma$.

In this case, the linear control analysis techniques cannot be applied directly to set the value of the FLL gain γ due to the frequency adaptation loop being nonlinear. As seen in [6], the averaged dynamics of the FLL with $\omega' \approx \omega$ can be described by:

$$\dot{\bar{\omega}}' = -\frac{\gamma V^2}{k\omega'}(\bar{\omega}' - \omega) \tag{4}$$

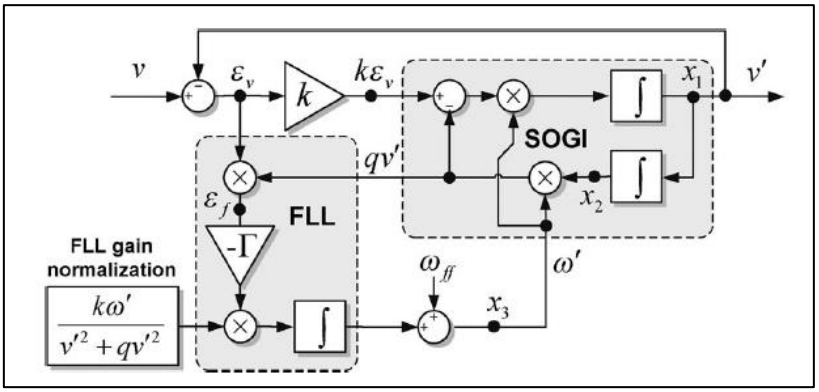


FIGURE 3. SOGI-FLL with feedback-based FLL gain normalization [1].

According to the equation (4) the value of γ can be normalized:

$$\gamma = -\frac{k\omega'}{V^2}\Gamma \tag{5}$$

to obtain the feedback-based linearized system shown in Fig. 4. This system does not depend on either the grid variables or the SOGI-QSG gain.
The FLL gain normalization block, shown in Fig. 3, computes the SOGI control parameter k as well as the output variables ω' and $V^2 = v'^2_2 + qv'^2$ to achieve the response linearization of the FLL. The time constant Γ is the parameter for setting the dynamics of the frequency estimation.

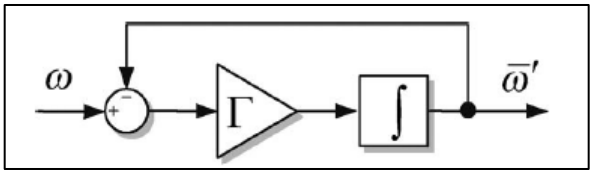


FIGURE 4. Simplified frequency adaptation system of the FLL [1].

The transfer function of the first-order frequency adaptation loop in Figure 4 is given by:

$$\frac{\overline{\omega}'}{\omega} = \frac{\Gamma}{s+\Gamma} \tag{6}$$

The settling time is highly dependent on the gain parameter Γ and can be approximated by:

$$t_{s(FLL)} \approx \frac{5}{\Gamma} \tag{7}$$

FIGURE shows the time response of SOGI-FLL with k = 2 and Γ = 50 when the frequency of the input grid signal suddenly varies from 50 to 45 Hz. As the figure shows, the detected frequency fits a first-order exponential response. The settling time is 100 ms, which matches with the calculation in equation (7) [1].

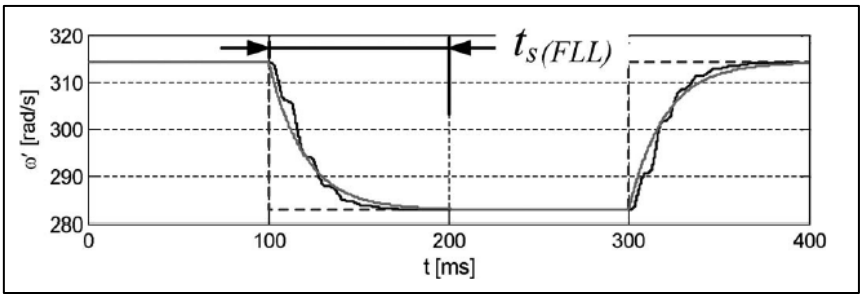


FIGURE 5. Time response of FLL in presence of a frequency step [1].

One important thing to conclude, regarding the frequency estimation, is that the FLL system estimates the frequency error, $\Delta\varepsilon_f$. Let us suppose step change frequency $\Delta\varepsilon_f$ at the input of the FLL. Laplace transformation of the step change frequency is given with $\frac{\Delta\varepsilon f}{s}$. Using the final value theorem (8), it is easy to prove that in the steady-state, FLL estimates step frequency change $\Delta\varepsilon_f$ (9).

$$\lim_{t\to\infty} f(t) = \lim_{s\to 0} sF(s) \tag{8}$$

$$\lim_{s\to 0}\left(s * \frac{\Gamma}{s+\Gamma}(s) * \frac{\Delta\varepsilon_f}{s}\right) = \Delta\varepsilon_r \tag{9}$$

By plugging s to be zero, the expression goes to $\Delta\varepsilon_f$, proving and simplifying the conclusion regarding the frequency estimation.

Proposed power calculation method

Power calculation, for the measurement system proposed in this paper, is performed using addition theorems from trigonometry. Since the SOGI-FLL subsystem gives two output components of voltage/current in quadrature (phase-shifted for $\pi/2$), *v'* and *qv'* or *i'* and *qi'* (Figure 2) respectively, they can be used to calculate powers rather simplistically. Considering that values *v'*, *qv'*, *i'* and *qi'* are $Vcos\varphi_u$, $Vsin\varphi_u$, $Icos\varphi_i$ and $Isin\varphi_i$, these fit perfectly into trigonometric addition theorems.

$$Vsin\varphi_u * Isin\varphi_i + Vcos\varphi_u * Icos\varphi_i = VI(cos\varphi_u cos\varphi_i + sin\varphi_u sin\varphi_i) \tag{10}$$

From equation (10), the right side represents the cosine addition theorem cos(α-β). By applying the mentioned theorem, the following is true.

$$P = Vmax * Imax(cos\varphi)/2 \tag{11}$$

Since most power measurement systems consider RMS values, equation (11) can be rewritten in form:

$$P = \frac{V_{max}the}{\sqrt{2}} * \frac{I_{max}}{\sqrt{2}}(cos\varphi) \tag{12}$$

or:

$$P = V_{rmsilarly} * I_{rms}(cos\varphi) \tag{13}$$

Sim , reactive power can be calculated.

$$Q = V_{rms} * I_{rms}(sin\varphi) \tag{14}$$

Equation (14) is obtained from the sine addition theorem, which is simply achieved by:

$$Vsin\varphi_u * Isin\varphi_i - Vcos\varphi_u * Icos\varphi_i = VI(cos\varphi_u sin\varphi_i - sin\varphi_u cos\varphi_i) \tag{15}$$

According to equations (13) and (14), apparent power (S) can be calculated as follows:

$$S = \sqrt{P^2 + Q^2} = \sqrt{\left(V_{rms} * I_{rms}(cos\varphi)\right)^2 + \left(V_{rms} * I_{rms}(sin\varphi)\right)^2} \tag{16}$$

When solved:

$$S = V_{rms} * I_{rms} \tag{17}$$

From equations (13) and (17), power factor ($cos\varphi$) can be calculated as follows:

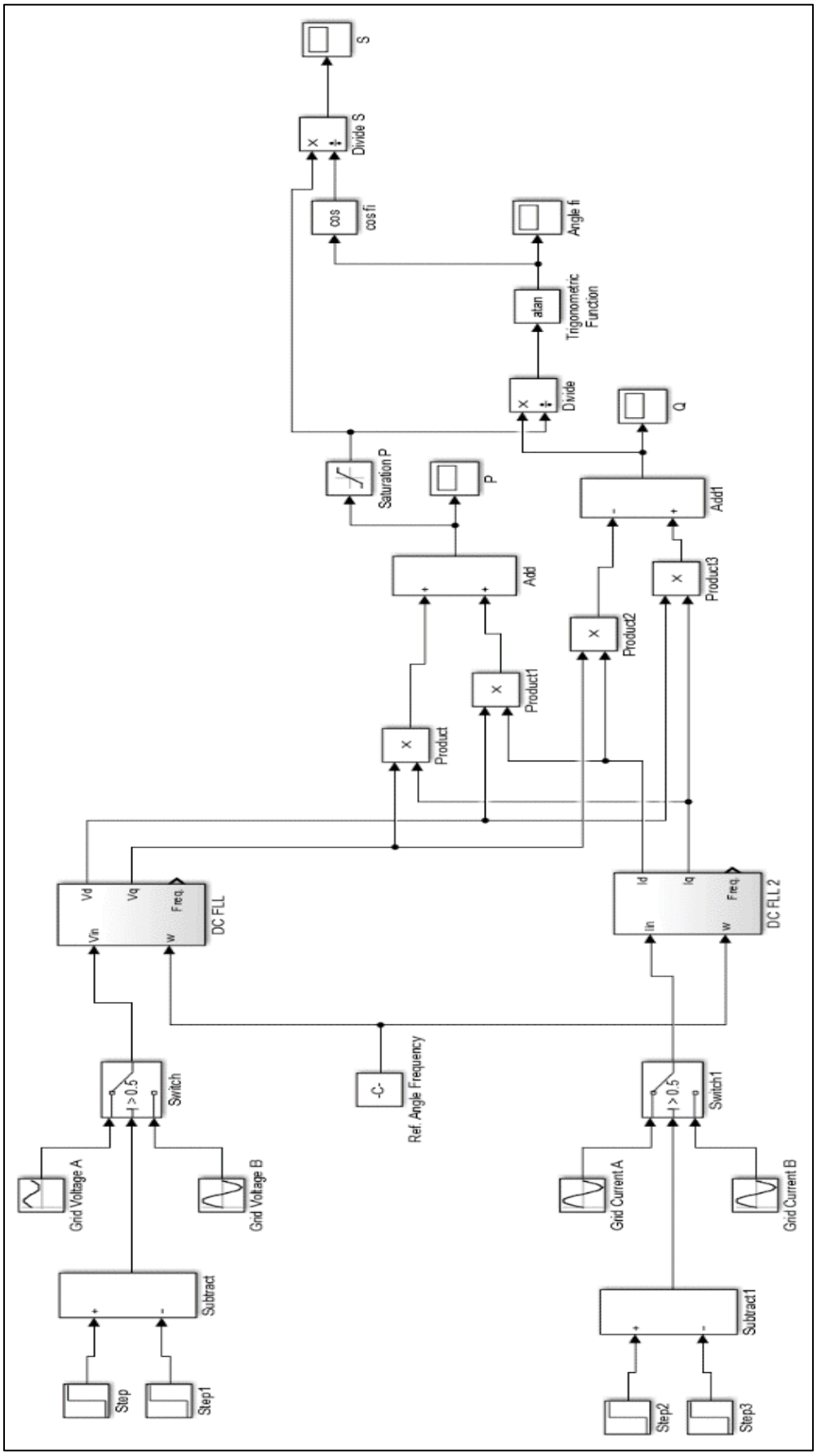
$$cos\varphi = \frac{P}{S} \tag{18}$$

With equation (18), all parameters for power calculation are complete.

4. Results and discussion

Performance of the suggested power measuring algorithm, for calculating the active (P), reactive (Q) and apparent (S) powers, were tested in Matlab/Simulink. The model of the suggested algorithm is shown in the following picture.

FIGURE 6 (on the next page). FLL Power Measuring System.


S
Divide S
cos
cos fi
Angle fi
atan
Trigonometric Function
Divide
Q
Saturation P
P
Add1
Add
Product3
Product2
Product1
Product
Vd
Vq
Freq.
Vin
w
DC FLL
Id
Iq
Freq.
In
w
DC FLL 2
Switch
> 0.5
Ref Angle Frequency
Switch1
> 0.5
Grid Voltage A
Grid Voltage B
Grid Current A
Grid Current B
Subtract
Subtract1
Step
Step1
Step2
Step3

This document contains an analysis of several scenarios, which are presented to showcase the capabilities of the FLL system used in this project. Some scenarios may contain multiple issues whereas others will focus on a single fault. Referent input values, for all scenarios, are:

- V = 220 [V]
- I = 5 [A]
- ω = 2*π*50 [rad/sec]
- V – I phase difference scenarios]

A. Scenario 1 Showcase

This scenario intends to show case how the system behaves when the gain of the FLL loop (Γ) is changed. Value of Γ had been optimally set to 50 in all upcoming scenarios. The system will undergo frequency oscillations, albeit, at different values of gain Γ. Additionally, the system will experience phase adjustment.

Figures 7 and 8 show the power measurement for the system when gain Γ is brought down from 50 to 10. By decreasing the value of Γ, the system achieved a stable output in a period, which can be seen from equation (7). The conclusion from this examination is that the system output behaves more oscillatory, as opposed to the optimum value of Γ = 50.

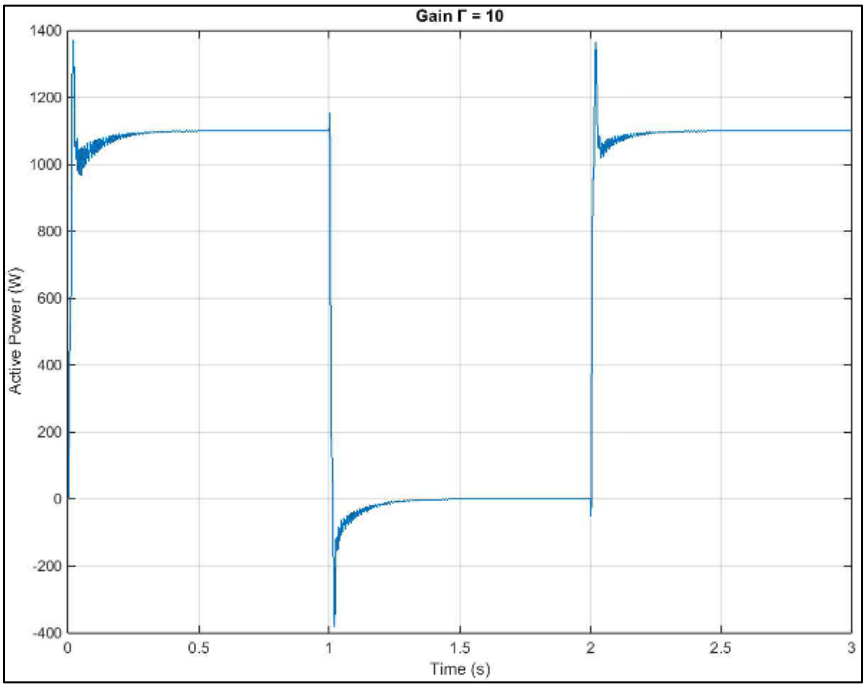


FIGURE 7. Active power for scenario 1 (Γ = 10).

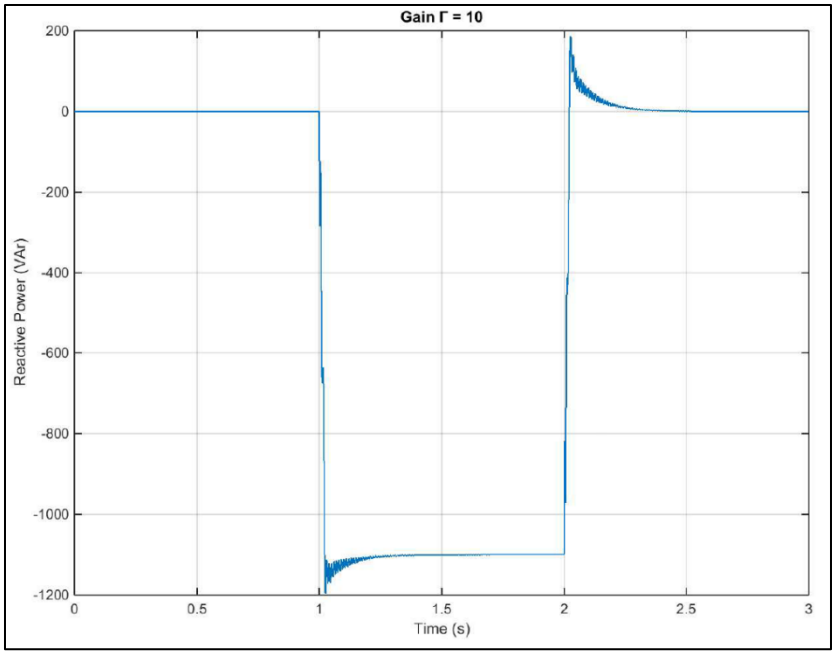


FIGURE 8. Reactive power for scenario 1 ($\Gamma = 10$).

A slower response may not necessarily be a bad thing. Some systems demand lower responses, and in such cases, a lower value of gain Γ may be the optimal solution.

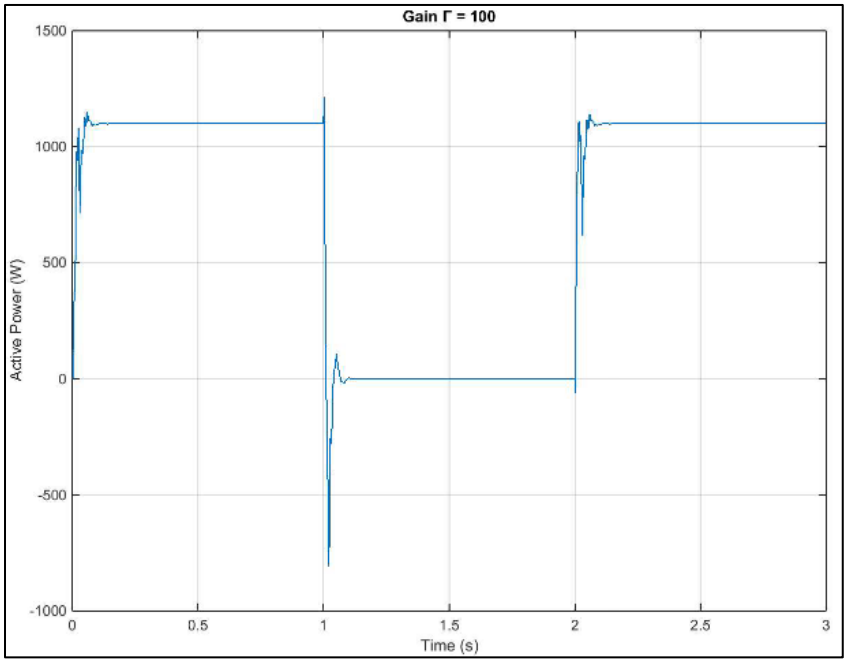


FIGURE 9. Active power for scenario 1 ($\Gamma = 100$).

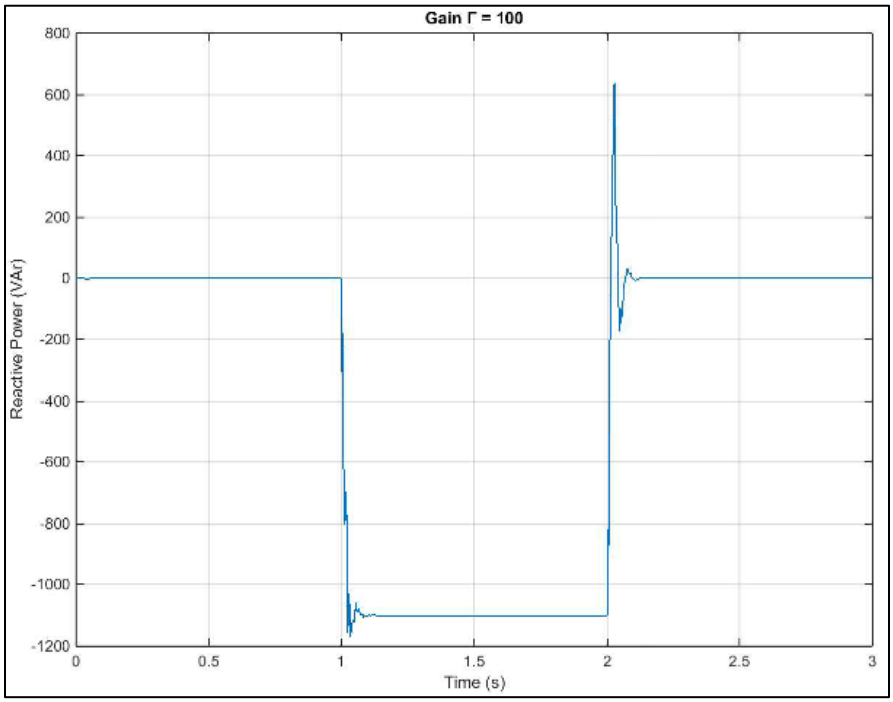


FIGURE 10. Reactive power for scenario 1 ($\Gamma = 100$).

Following the same equation (7), a higher value of gain Γ ought to cause a faster system response. While this may be true in theory, Figures 9 and 10 show the actual stability. The time it takes to stabilise the output is certainly much shorter than in Figure 7. However, the transition process experiences a much greater overshoot, as well as the unpredictable oscillations, before the signal itself is stabilised. This shows that the gain Γ cannot be infinitely increased. The increase depends on the measured system and the need for real-time information. As it was previously mentioned, the optimal value of Γ in this paper was calculated at $\Gamma = 50$.

B. Scenario 2 – Referent Measurement

In scenario 2, the regular operation mode is considered. In this case, there are no faults on the input, as this measurement serves as a reference to all the following scenarios.
Figure 11 shows the graph of active power (*P*). The measured value of *P* is 1100 W. Considering that no changes were present (in comparison to referent values) on inputs, this calculation is a simple multiplication of voltage and current.

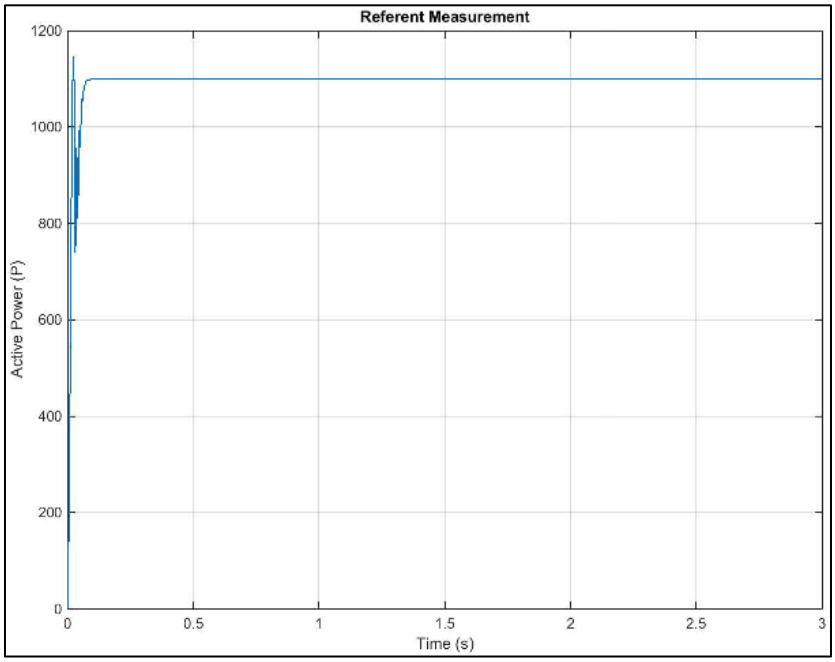


FIGURE 11. Active power for scenario 2.

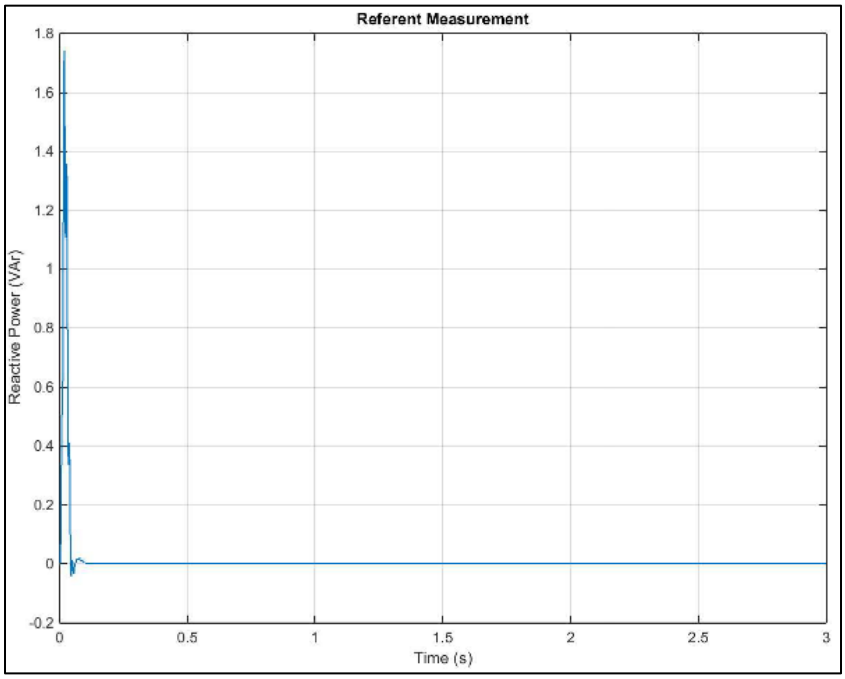


FIGURE 12. Reactive power for scenario 2.

Figure 12 follows the same principle. Since current and voltage are in phase, the value for reactive power (*Q*) is zero.

C. Scenario 3 – Phase Adjustment

Scenario 3 examines the case where the phase of one input (in this case, this is the voltage phase) is shifted during the operation of the system. At t = 1 s, the voltage phase is changed from 0 to $\pi/2$. At t = 2 s, this change is reset, to clarify the system behavior.

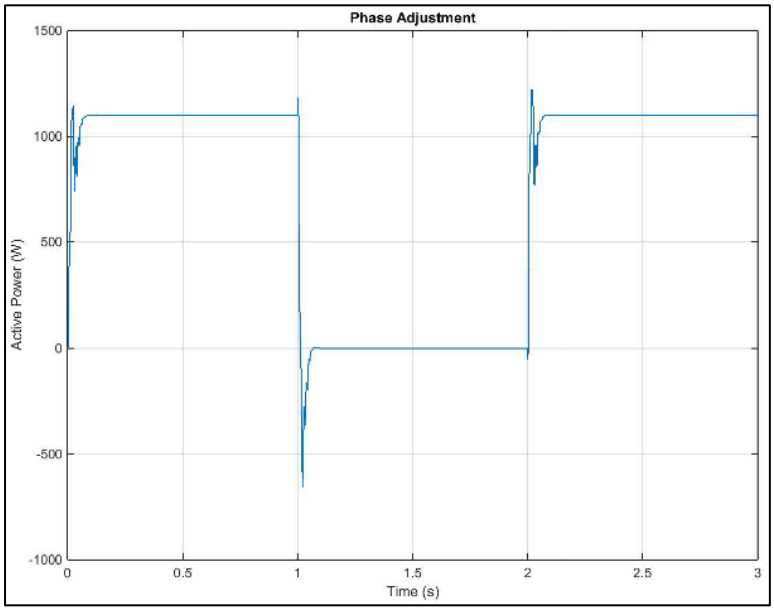


FIGURE 13. Active power for scenario 3.

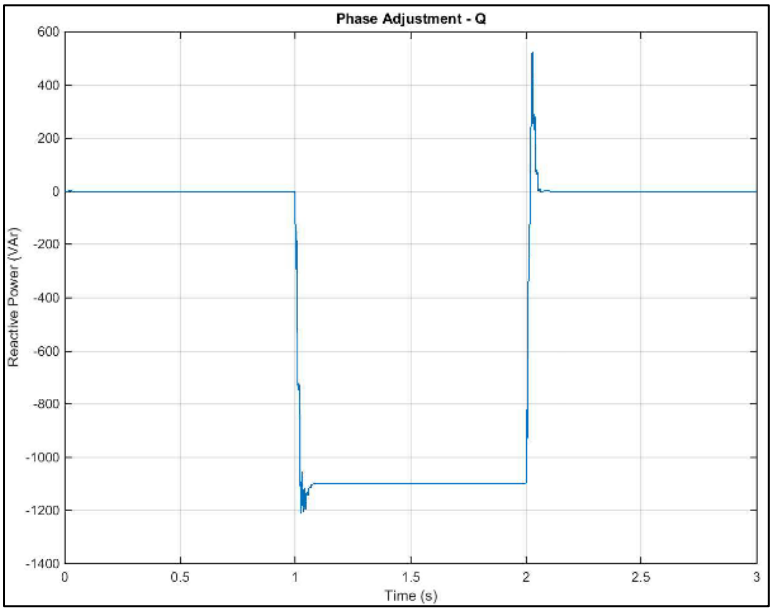


FIGURE 14. Reactive power for scenario 3.

Figure 13 shows active power when the voltage phase is adjusted. For the first second, both

voltage and current are in phase. During this time interval, active power is equal to the referent power in scenario 2, being P = 1100 W. At t = 1 s, the voltage phase switches from 0 to $\pi/2$. During this time interval, until t = 2 s, active power is zero, due to voltage and current having a phase difference of $\pi/2$, leading to ($\cos(\pi/2) = 0$).
Figure 14 shows reactive power for scenario 3. Up until t = 1 s, reactive power is Q = 0 VAr, due to voltage and current being in phase. At t = 1 s, the voltage phase changes to $\pi/2$, leading to reactive power Q = -1100 VAr. Since the voltage phase is the one changing, the angle is $\varphi = -\pi/2$.

$Q = 220\ V * 5\ A * \sin(-\pi/2) = -1100\ VAr$

At t = 2 s and after, the voltage phase goes back to zero, therefore the reactive power goes back to Q = 0 VAr.

D. Scenario 4 – Frequency Oscillation

Frequency values for voltage and current have been known to oscillate in known systems. Due to this fact, the presented FLL system will showcase its behavior when such a change occurs.

For this scenario, a current phase has been set to $\pi/2$. At t = 1 s, both voltage and current go through a step frequency change from 50 Hz to 55 Hz. At t = 2 s, changed values revert to initial ones.

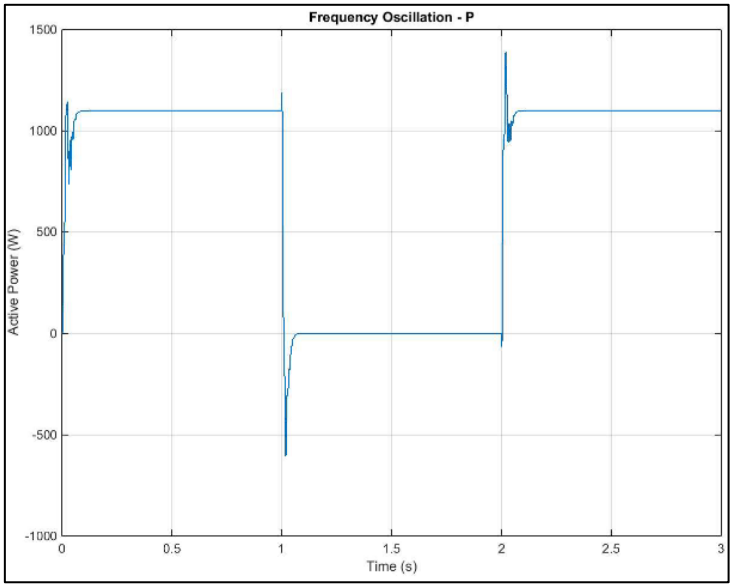


FIGURE 15. Active power for scenario 4.

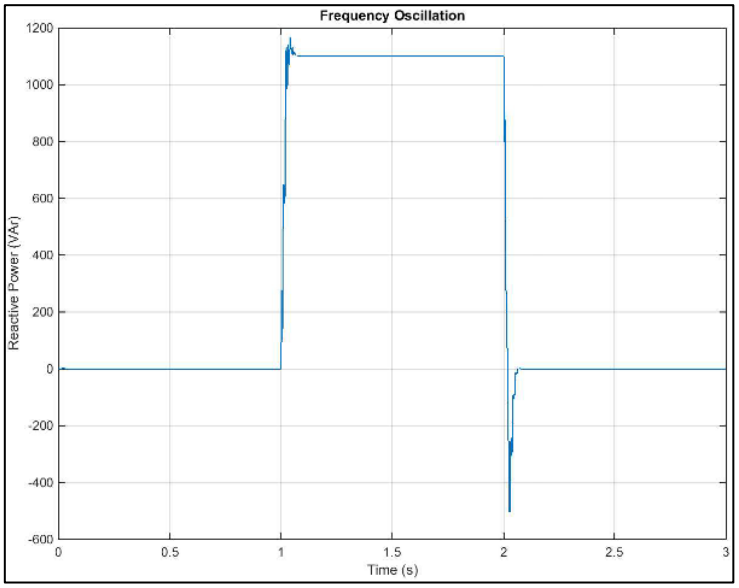


FIGURE 16. Reactive power for scenario 4.

Figure 15 shows the measured active power as frequency oscillations occur. At t = 1 s, frequency values for voltage and current change, and the system clearly shows this through the oscillation of the active power, on the graph. This oscillatory behavior can be affected through the value of Γ. Value of Γ used in this project is deemed to be optimal, however, that can be further debated. Regardless, in a short amount of time, the system recovers, and accurately proceeds to perform the measurement, showing correct values of active power. Similar oscillatory behavior repeats once the measured system restores the initial value of frequency (50 Hz). This measurement is more than satisfactory.
Figure 16 proves that the same logic applies to reactive power, and that the algorithm is capable of solving multiple issues at once.

5. Conclusion

In conclusion, this paper elaborated on the usefulness of the FLL in power measuring systems. While the FLL may be used in different systems, this paper proves the versatility and the capability of the FLL to perform power measuring. The FLL was successfully performed through various problematic scenarios, such as phase adjustment and frequency oscillation.

The power measuring design, shown in the project, greatly reduces the demand for DSP systems, which need to perform integrals and derivatives to calculate powers. Instead, by utilizing the output of the SOGI, simple trigonometry is enough to perform accurate and fast power calculations.

Additionally, the currently present power measuring FLL system can be further improved. It is a menial task to add apparent power and power factor calculations. Another issue that the FLL can resolve is the issue of the DC component which causes major issues in ordinary power measuring systems. However, this topic will be further expanded on in future works. The presented system performed marvelously, completely fulfilling expected results.

6. References

[1] C. f. O. P. o. t. E. Communities, "Towards Smart Power Networks—Lessons Learned From European Research FP5 Projects," 2005.

[2] Blanchard, Phase-Locked Loops - Application to Coherent Reciever Design.

[3] Ciobotaru, M., Teodorescu, R., & Blaabjerg, F. (2006). A New Single-Phase PLL Structure Based on Second Order Generalized Integrator. *37th IEEE Power Electronics Specialists Conference*, 1–6. https://doi.org/10.1109/PESC.2006.1711988

[4] Freijedo, F. D., Doval-Gandoy, J., Lopez, O., & Acha, E. (2009). Tuning of Phase-Locked Loops for Power Converters Under Distorted Utility Conditions. *IEEE Transactions on Industry Applications*, *45*(6), 2039–2047. https://doi.org/10.1109/TIA.2009.2031790

[5] Golestan, S., Guerrero, J. M., Vasquez, Juan. C., Abusorrah, A. M., & Al-Turki, Y. (2018). Modeling, Tuning, and Performance Comparison of Second-Order-Generalized-Integrator-Based FLLs. *IEEE Transactions on Power Electronics*, *33*(12), 10229–10239. https://doi.org/10.1109/TPEL.2018.2808246

[6] Luna, A., Rocabert, J., Candela, J. I., Hermoso, J. R., Teodorescu, R., Blaabjerg, F., & Rodriguez, P. (2015). Grid Voltage Synchronization for Distributed Generation Systems Under Grid Fault Conditions. *IEEE Transactions on Industry Applications*, *51*(4), 3414–3425. https://doi.org/10.1109/TIA.2015.2391436

[7] Rioual, P., Pouliquen, H., & Louis, J.-P. (1996). Regulation of a PWM rectifier in the unbalanced network state using a generalized model. *IEEE Transactions on Power Electronics*, *11*(3), 495–502. https://doi.org/10.1109/63.491644

[8] Robles, E., Ceballos, S., Pou, J., Martín, J. L., Zaragoza, J., & Ibañez, P. (2010). Variable-Frequency Grid-Sequence Detector Based on a Quasi-Ideal Low-Pass Filter Stage and a Phase-Locked Loop. *IEEE Transactions on Power Electronics*, *25*(10), 2552–2563. https://doi.org/10.1109/TPEL.2010.2050492

[9] Rodriguez, P., Luna, A., Candela, I., Mujal, R., Teodorescu, R., & Blaabjerg, F. (2011). Multiresonant Frequency-Locked Loop for Grid Synchronization of Power Converters Under Distorted Grid Conditions. *IEEE Transactions on Industrial Electronics*, *58*(1), 127–138. https://doi.org/10.1109/TIE.2010.2042420

[10] Rodriguez, P., Luna, A., Candela, I., Teodorescu, R., & Blaabjerg, F. (2008). Grid synchronization of power converters using multiple second order generalized integrators. *2008 34th Annual Conference of IEEE Industrial Electronics*, 755–760. https://doi.org/10.1109/IECON.2008.4758048

[11] Wu, F., Sun, D., Zhang, L., & Duan, J. (2015). Influence of Plugging DC Offset Estimation Integrator in Single-Phase EPLL and Alternative Scheme to Eliminate Effect of Input DC Offset and Harmonics. *IEEE Transactions on Industrial Electronics*, *62*(8), 4823–4831. https://doi.org/10.1109/TIE.2015.2405496

[12] Yu, B. (2018). An Improved Frequency Measurement Method from the Digital PLL Structure for Single-Phase Grid-Connected PV Applications. *Electronics*, *7*(8), 150. https://doi.org/10.3390/electronics7080150

Prediction of Solved Homicides Using a Classification Method

Lamija Zukic*, Samed Jukic*

*International Burch University
lamija.zukic@stu.edu.ibu.ba
samed.jukic@ibu.edu.ba

Original research

Abstract: *Homicide rates are still high in the world and they are the worst crime in human existence. Despite all the technological advances and usage of information by various agencies, the number of homicides is not decreasing. Homicide prediction in certain countries should notably be the number one priority, which can help the government to easily identify the kind of profile they are looking for, or even help them prevent those cases. This paper compares different Machine Learning Techniques classifications of homicide prediction. Random Forest (RF), Random Tree, J48, Naive Bayes and k-Nearest-Neighbor (KNN) were tested to determine which method provides the best results in homicide prediction classification. The results of sample accuracy for all algorithms were around 99%, which clearly shows that all algorithms give great results. However, J48 is the best technique applied on the dataset, as it classified all instances correctly.*

Keywords: Classification, data analysis, homicide, machine learning, prediction.

1. Introduction

Homicide continues to prevail in most news. Unfortunately, every evening news ends up broadcasting at least one homicide within the country's region. This has been an issue and still is in most countries, but especially in the United States of America (USA). It is widely known that USA government agencies have Crime Departments that are advanced in the fields of data gathering and data processing for various needs. However, it seems that the homicide rate in the USA is still high, and it is a big problem as homicide is the most violent form of crime. Term homicide clearance rate refers to the percent of homicides that lead to arrest or charge, meaning that the case can be declared as solved.

On the other hand, uncleared homicide means that the suspect is not found due to the lack of data gathered by the police. This can lead to a serious problem, as the perpetrator is freely moving, and can result in another homicide as revenge. Homicide clearance rates currently range up to 65% in the USA [1], 95% in Japan [1] and 75% in Canada [2]. With evolving technologies, nowadays many tools are used for analytics and predictions in crimes. Having a huge amount of raw and meaningful data creates many opportunities for getting the desired output that can help the government make accurate decisions. In the case of homicide, that should be the number one priority, crime agencies should use that data to make predictions to prevent homicide crime. Various machines are applied for different reasons and different outputs.

This research paper uses a USA homicide dataset to predict homicide clearance, in terms of whether the crime is solved or not. Before ML models are applied, data is preprocessed and cleaned. Different machine learning methods are applied to the data. Random Forest (RF), Random Tree, C4.5, Naive Bayes and k-Nearest-Neighbor (KNN) were compared based on different performance evaluation criteria.

The organization of the paper is as follows. Section 2 presents literature background work on the prediction of homicide, whereas Section 3 describes the Homicide dataset and ML techniques applied. Section 4 presents the experimental results. Finally, Section 5 concludes the paper.

2. Literature Review

Few research papers analysis of homicide clearance data in certain countries and the factors that lead to the decision whether the crime is solved or not. A paper *The Value of Life in Deaths provide* used multiple regression to determine there were extralegal factors, particularly racial and gender, that affected clearance rates. The results showed that homicide clearance varied by several extralegal factors. For instance, cases involving non-white victims or older victims were less likely to be solved, thus white victim homicides had a 42 percent greater chance of being solved than non-white cases. Also, cases that involved younger children were more likely to be solved than cases with older victims. Multiple regression used in this paper resulted in binary outcomes that predicted the chance of homicide clearance, and it determined which homicides were most likely to be cleared. The model proved to be very successful [3].

Another paper used a sophisticated statistical approach multilevel latent class analysis [MLCA], which concluded that more stranger homicides were not solved, which might have been due to missing data on the victim/perpetrator relationship. Also, robberies or similar felonies that resulted in homicides may be mistakenly characterized as stranger killings [4]. McClendon and Meghanathan implemented three algorithms, Linear Regression, Additive Regression, and Decision Stump (DS) using the same finite set of features on crime unnormalized dataset to identify violent crime patterns from two datasets [5]. Regression proved to be the best method for predicting the crime data based on the training set input. The algorithm that had the greatest

correlation coefficient value and generated the lowest error values is the linear regression algorithm. The DS algorithm proved to be the least accurate one.

Another paper performed ML algorithms in crime prediction in Canada. Kim and Joshi used classification methods to identify patterns and create predictions. K-nearest neighbor (KNN) and Decision Tree algorithms were implemented to analyze the crime dataset. Their dataset consisted of more than 500,000 records with an accuracy between 39% and 44%. The accuracy turned out to be below as a prediction model, however, the accuracy can be increased by tuning the algorithms and crime data for specific applications [6]. Rolf Loeber & Lia Ahonen examined the Pittsburgh Youth Study, which began in 1987. The study was conducted among boys from public schools who were randomly selected from 1, 4, and 7 grades, respectively. To identify the number of high-risk males, they used a so-called screening assessment, which consists of collecting information from participants, their parents, and teachers. For the next assessment, 30% of the most antisocial boys were chosen, with 30% of boys randomly selected from the remaining 70%. In the following assessment, selected boys had carried out face-to-face conversations at half-yearly intervals throughout approximately, the next 10 years [10]. According to Rolf Loeber & Lia Ahonen, the study is uniquely created to investigate individuals' delinquency and substance use, as well as their continuation and renouncement from such behavior. The information collected up to 2009, 37 males were convicted for homicide between the ages 15 and 29, while in total 39 males were victims of homicide. The average age of perpetual is 19.7, while the average age of the victim is 22.7. An interesting data gathered from the study is that 32 out of 37 convicted homicide offenders were African American, and 37 out of 39 are homicide victims. Another important piece of data pointed out in this study is the usage of guns as the main weapons. To sum up, these homicides were mostly carried out by African American males that involved guns, drugs, and gangs. There were few anomalies, but such were excluded from the result analysis. In this study, the authors explain how three different types of predictors need to be considered: factors in family and neighborhood environment, early behavioral factors such as conduct problems, and early offenses such as self-report and arrest. The study also identifies the prediction of homicide victims, using regression analysis. Same characteristics and predictors are shared between victims and offenders [10].

This study worked on a detailed and thorough analysis of a small and selected number of participants, which involved yearly face-to-face conversations throughout the years, where people were able to follow their life and their behavior on a certain basis. Predictions in terms of predictors were identified, as to what factors in one's life can lead up to homicide and victim.

3. Dataset and Machine Learning Techniques

3.1. Dataset

The homicide database in the United States, compiled and gathered by the Murder Accountability Project is made available to the world. The dataset includes murders from the FBI's Supplementary Homicide Report from 1980 to 2014, rounding up to almost 35 years, on more than 600,000 murders. This research paper uses homicide records from 2010 to 2014 as those are the most recent data available to the public. The dataset contains 20 relevant features, such as information about homicide time and place, victim and perpetrator details, the relationship between the two, and the weapon used. Dataset has been preprocessed to clear noisy data to provide more meaningful results. Python language has been used for data cleaning. Crime in the US has been a big topic for decades between the crime agencies and politics. As each country is fighting to drop its crime rate, so is the US. Having an accurate database of such information can be helpful in the future once applying a proper ML model thus analyzing the results.

3.2. Data Preprocessing

Python language continues to be a very popular and well-maintained open-source programming language in Data Science. Before ML models were performed, the dataset went through data preprocessing in Python. A new data frame has been created and later used for machine learning. Some features were not relevant, so they were excluded from the dataset. Due to the very large dataset only records for the years 2010 through 2014 were taken into account, as they are the most recent public data. Records where the feature Relationship is Unknown and where MurderSolved has value Yes, were removed from the dataset as well. These records had most of the features classified as unknown, and such data would not contribute to the research. However, this significantly reduced time and CPU memory for the model to build. Also, certain features, such as victim age and perpeterur age had few values identified as outliers and they made no sense. This problem would impact the analysis, so methods such as aggregation and mean replacement were used to overcome this problem.

3.3. Machine Learning Techniques

Classification is one of the most common applications in Machine Learning. It identifies and discovers patterns, and later on sort the data into groups based on its similarities. Classification methods build a model that predicts future outcomes using predefined classes which are based on certain criteria [12].

3.4 J48

J48 is an open-source Java implementation of the C4.5 decision tree algorithm. J48 as a decision tree classifier has additional features for missing data, continuous attribute value ranges, pruning of decision trees, rule derivation, etc. J48 uses a predictive ML model that calculates the final value from the dataset. Its structure consists of root nodes, intermediate nodes, and leaf nodes, where each node consists of a decision, leading to the final result. To calculate which attribute is the best option for splitting the tree, we use the splitting criterion [7].

3.5 Random Forest (RF)

Random Forest is a method that operates by creating multiple decision trees during its first phase or training phase. The decision of the majority of the trees is chosen by the RF as its final result. The first benefit of RF is the reduction of overfitting, as we use multiple trees, meaning that the data is fit so close. RF runs efficiently on large data, thus it produces highly accurate predictions. Also, it estimates missing data, thus maintaining accuracy when a large proportion of data is missing [8].

3.6 Naive Bayes Classifier

Naive Bayes is another classification algorithm that assumes the input values are nominal, even though numerical inputs are also supported once applied. Naive Bayes uses an implementation of the Bayes Theorem which is based on the principles of conditional probability, as each class is obtained from the training set and is adopted as independent variables. It later predicts the class with the highest probability. This algorithm has proved to be very effective (fast and easy to calculate), despite the impractical assumption where the variables are expected to be independent [7].

3.7 K-Nearest-Neighbor (KNN)

KNN is also called instance-based learning. KNN algorithm supports both classification and regression methods. It is a simple algorithm that locates or classifies a new instance closest or the most similar to the training patterns. To make a prediction, it takes the mode (most

common class) of the training pattern to find the k, most similar instance. KNN method produces a linear decision boundary [7].

4. Results and Discussion

This section presents all of the results from the implementations of the J48, Random Forest, Naive Bayes, and KNN algorithms. The algorithms were run to predict the feature CrimeSolved. The algorithm that gives the lowest error values for prediction and its accuracy is highlighted in the results presented in Tables 1 and 2.

TABLE 1. Algorithm Results.

	Sample Accuracy	**ROC Area**	**F-Measure**
J48	100%	1	1
RF	99.9%	1	1
Naive Bayes	99.9%	1	1
KNN	99.8%	0.99	0.91

TABLE 2. Confusion Matrix.

Algorithm	**Confusion Matrix**	
	Correct	Incorrect
J48	34914	0
Random Forest	34904	10
Naive Bayes	34904	10
KNN	34878	36

For all these algorithms, cross-validation of 10 folds was employed as a parameter before the model building. Cross-validation is a standard evaluation technique, which systematically runs a repeated percentage split. As a result, it gives 10 evaluation results, which are later averaged. The sole purpose of this is to give better results since the dataset is relatively large. The total number of instances was 34914. The percentage of correctly classified instances is usually called sample accuracy. All four classifiers have been implemented and tested in software package WEKA using default parameters.

Looking at the table, the overall accuracy of the algorithms provides very similar and r. The algorithm that had the greatest sample accuracy was J48 (100%) among other three algorithms, with a decision tree as an output. KNN algorithm was the least accurate model (99.8%) with the

most incorrectly classified instances (36). The model with zero incorrectly classified instances was J48.
Another preferred measure is F-Measure. F-Measure is a combined measure for precision and recall calculated as follows:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

The ROC area for all four algorithms is high, meaning that if we were given an item from both classes, this would be a percentage of randomly putting them correctly. Given the classification task for this particular dataset and the features that have been provided, J48 and Random Forest algorithm are the most accurate of the four.
Thomas Hargrove, a founder of a nonprofit organization for crimes, developed an algorithm for detecting serial killers and their activities. The algorithm is based on clustering methods where murders of women within a close area and similar weapons are used. The algorithm organizes homicide into clusters based on victims' data, such as gender, geographic location, and weapon used, and outputs murder groups with low homicide clearance rates [9]. According to the source, the algorithm proved to be successful in catching serial killers, one in particular.

Y. Rayhan and T. Hashem developed a prediction model using Spatial-temporal systems, which describes a phenomenon in a certain location and time. The model is called AIST a.k.a. Attention Based Interpretable Spatio Temporal Network for crime prediction, and it uses historical data (past crimes), external features (e.g., traffic, point of interest), and recurring trends of crime. The model proved to be very accurate and reliable using real data. The paper compared this method to other methods such as decision tree and RNN, where the AIST method outperformed most algorithms, looking at the evaluation criteria [11].

5. Conclusion

We observe all four algorithms to be very effective and accurate in predicting the homicide clearance data based on the training set input. Looking at the confusion matrix, we can have a deeper insight into the accuracy of models. The J48 algorithm classified correctly all instances in the dataset. The reason for such accuracy might certainly be due to the data preprocessing. Certain data was transformed to bring it to a state where the algorithm can easily parse it, meaning that the data can be easily interpreted by the algorithm. Data preprocessing is a crucial part, as it directly impacts the accuracy rate.

After an analysis of all the homicides that happened between 2010 and 2014, it is concluded that most homicides are solved while a small percentage of homicides are unsolved. As the idea is to find patterns between homicides that are unsolved, the results showed interesting information. Algorithms performed provided high accuracy in classifying unsolved homicides, and the following text will explain what attributes contribute to these results. The homicide rate is slightly decreasing. This is due to evolving technologies in the world, where the USA is investing huge amounts of money, especially in government and security facilities. Tools such as CCTVs and systems can help agencies solve crimes in much easier and more convenient ways.
Predicting homicide clearance can be a crucial contribution when solving a case. Knowing a potential outcome, whether a case is solved or not makes solving a crime case easier and is a helpful tool for the government, police, and investigators.

6. References

[1] Roberts, A., (2008). *Explaining Differences in Homicide Clearance Rates Between Japan and the United States*.

[2] Mahony, TH, Turner, J (2012) *Police reported clearance rates in Canada, 2010*. Juristat article. Cat no 85-002-X. Ottawa: Statistics Canada.

[3] Lee, C. (2005). *The value of life in death: Multiple regression and event history analyses of homicide clearance in Los Angeles County*. Journal of Criminal Justice 33.

[4] Flewelling, R. L., & Williams, K. R. (1999). *Categorizing homicides: The use of disaggregated data in homicide research*. In M. D. Smith & M. A. Zahn (Eds.), Homicide: A sourcebook of social research (pp. 96 – 106).

[5] McClendon, L. and Meghanathan, N., (2015). *Using Machine Learning Algorithms to Analyze Crime Data*. ResearchGate OI: 10.5121/mlaij.2015.2101

[6] Kim, S., Joshi, P., Kalsi, P. and Taheri, P., (2018). *Crime Analysis Through Machine Learning*. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCONalog).

[7] Brownlee, J., (2016). *How To Use Classification Machine Learning Algorithms in Weka*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/>

[8] Simplilearn. 2018. Random Forest Algorithm - Random Forest Explained | Random Forest in Machine Learning | Simplilearn. [online] Available at: <https://www.youtube.com/watch?v=eM4uJ6XGnSM>

[9] Kolker, R., (2017). Bloomberg - *Serial killers should fear this algorithm*.[online] Bloomberg.com. Available at: https://www.bloomberg.com/news/features/2017-02-08/serial-killers-should-fear-this-algorithm

[10] Rolf Loeber & Lia Ahonen (2013). Journal of Youth and Adolescence - *Invited Address: Street Killings: Prediction of Homicide Offenders and Their Victims*.

[11] Y. Rayhan and T. Hashem (2020). arXiv.org - AIST: An Interpretable Attention-based Deep learning Model for Crime Prediction

[12] J. Brownlee (2020). https://machinelearningmastery.com/ - 4 Types of Classification Tasks in Machine Learning

Machine Learning-Based Gene Clustering on Brain Cancer Using K-Means and Hierarchical Clustering Methods

Fatih Yilmaz*, Samed Jukic*
*International Burch University
fatih.yilmaz@stu.ibu.edu.ba
samed.jukic@ibu.edu.ba

Original research

Abstract: *K-means and hierarchical clustering algorithms are employed to cluster genes according to the gene expression to determine the harming level of the genes in brain cancer. The gene expression data with a control group from The Cancer Genome Atlas database were used. The optimal cluster number for each clustering technique was obtained using the elbow method and dendrogram for K-means and hierarchical clustering methods respectively. We identified the ideal number of clusters as three and further classified them into seven groups. We observed that the second cluster contains over half the genes in healthy people and the cluster distribution of a healthy patient and a patient who died six months after being diagnosed with brain cancer is similar. Further analysis indicated that of all the time spent by patients after being diagnosed with brain cancer, group 0 has the highest percentage in one month after the diagnosis, while group -2 has the lowest percentage. Most genes shift their clusters when K-means and hierarchical clustering techniques we compared with the genes from the control and disease groups. The result of the measure of dissimilarity between the genes expression patterns indicates that the K-means technique outperforms the hierarchical technique with a higher rate of change in the cluster.*

Keywords***: Brain cancer, gene clustering, hierarchical clustering, K-means clustering, machine learning.***

1. Introduction

The brain is the most important part of the body because it controls the central nervous system of the human body. It takes charge of several actions of the body, thereby playing a key role in the human nervous system. One of the leading causes of death and a major obstacle in increasing life expectancy in the world is cancer [1]. A World Health Organization (WHO) report in 2019 indicated that cancer is the foremost or the second prominent cause of death before attaining the age of seventy in 112 countries of the world [2]. Brain cancer is one of several types of cancer diseases, and it develops in the brain [3]. Some warning signs of brain cancer amongst others include frequent headaches, speech changes, coordination problems, and memory loss. This type of cancer stays in the brain [4]. Categories of brain cancer are based on the development stage, origin, growth rate, and nature. The cancer of the brain can be of two types, either malignant or benign [4]. The malignant brain cancer cells attack nearby cells present in the spinal cord or brain; they have fast development rates. Benign brain cancer cells hardly attack the nearby healthy cells, they exhibit slow development rates and have distinctive borders. Brain cancer can be diagnosed either invasively or non-invasively. The invasive technique includes doing a small opening to collect cancer samples for necessary clinical tests, where the samples are subjected to microscopic examination to ascertain the malignancy. The non-invasive technique includes carrying out a physical examination of the brain and the entire body using imaging systems such as magnetic resonance imaging and computed tomography, which are quicker and harmless than the invasive approach. The imaging methods enable the radiologists to identify brain disorders, observe the pattern of development and assist in surgical preparation [5]. The introduction of dominant computing machines has led to the decrease in the cost of diagnostic hardware through the development and deployment of computer-aided tools for brain cancer diagnosis. These tools are projected to enhance radiologists' ability to accurately and consistently deliver quality diagnostic results. In this study, two machine learning models; hierarchical clustering and K means were developed to cluster the gene present in brain cancer. Clustering algorithms are the main computational tools employed in this study. Clustering analysis includes the process of data grouping into two or more clusters such that data points in the same cluster are like those in different clusters due to information retrieved through the data points [6]. Carrying out clustering analysis on groups of cancer samples having similar patterns can lead to the discovery of new cancer subtypes. Clustering analysis was first employed in the study, "Molecular Classification of Cancer" [7].

A. Clustering Techniques

Clustering techniques were used after the preceding procedures were completed properly. The following sections go over two different clustering techniques:

B. K-means Clustering

K-Means clustering aims to segregate data into groups, and usually, the grouping is occasionally characterized by the variable 'k'. The algorithm makes an effort to assign each information point to the variable 'k' groups available concerning the feature similarity. This method of clustering data is usually appropriate for use because it is relatively simple to implement the variables and at the same time generalize clusters of distinct shapes and sizes like the elliptical. It can also be used to scale large data sets, which saves time, and reduces the tediousness of the grouping. Additionally, K-Means clustering adapts quickly to new examples making it easy to understand and interpret hence it is useful because of its flexibility. Thus, in summary, K-means clustering is a technique used in objects in a procedure that minimizes their variation amongst them. Among the various types of existing K-mean algorithms, the one in

practice highlights the variation present in a group as the summation of the squared distances of Euclidean existing between each element of the group and centroid ad is given as

$$W(C_k) = \sum_{x_i \epsilon} (x_i - \mu k)^2 \quad (1)$$

In equation (1), x_i is the data that belongs to the C_k cluster, and μk is the mean value of the data given to the C_k cluster. [8]. Thus, the sum of all the K clusters divided by the summation of the Euclidean distances is then squared amongst the data and to the corresponding center.

C. Hierarchical Clustering

Ultimately, hierarchical clustering is another method of grouping data that seeks to set up a hierarchy of clusters that usually are comprised of two types; the Agglomerative type also known as the bottom-up approach ensures each observation begins in their clusters and a pair of clusters are brought together moving up the hierarchy. Divisive also known as the top-down approach ensures all the observations begin in one cluster. The splits are carried out recurrently as they move down the hierarchy. Notably, the results of this clustering method are always presented in the form of a dendrogram. Hierarchical clustering is beneficial because it is easy to implement and usually assures the best results in most of the areas it is applied to. Furthermore, there is always no specific information about the number of clusters required, making it suitable for every application [9].

2. Literature Review

The incessant process of unwanted death of cells apart from the production of new ones is controlled by the genes. The development of cancerous cells originates from uncontrolled cell growth. Medical imaging methods have helped health professionals and researchers to have a deeper view of the inner human body and carry out the analysis of this part without undergoing incisions. To accord proper cancer treatment, diagnosis, grade assessment, treatment response assessment, surgery planning, and patient prognosis are the key steps to follow. There are two brain imaging approaches, and they are functional and structural imaging [10]. The functional approach identifies the metabolic changes, cuts on an improved scale, and visualizes the activities of the brain. On the other hand, the structural approach includes several measures associated with brain cancer location, structure, injuries amongst others. MRI and CT are mostly used for brain cancer analysis which can capture various sections of the body without any surgery [11]. To further improve brain image analysis, several machine learning models have been used in describing brain tumors [12]. In the application of machine learning for brain image characterization, two important stages are involved: feature extraction and classification. The feature extraction stage involves a set of mathematical models that are built around some image characteristics such as contrast, texture, and brightness. To improve the perceived power of the model, several features accrued from various extraction models are joined together [13]. Brain images are classified and segmented using models such as Artificial Neural Network, K-Nearest Neighbors, Support Vector Machines, metaheuristic algorithms, region growing algorithms, and morphology amongst others.

In biomedicine, cluster analysis is a major data mining approach that is employed in data analysis processes. The hierarchical clustering technique is a classical clustering approach that has been broadly used in the field of biomedical. According to this study at BMC Bioinformatics, the major reason for using hierarchical clustering is its simplicity. It requires just one parameter, the number of clusters, and the availability of implementations as part of the software [14]. Another classical clustering algorithm is K-means, which is a method that requires cluster numbers to be given as input by the user. Generally, finding a suitable value for the cluster

numbers is a demanding task [15]. K-means has been identified as one of the best clustering algorithms used for analyzing cancer gene expression data [16], even with its non-deterministic nature issue, which is one of the main drawbacks of the K-means technique. Also, the K-means algorithm is relatively simple to implement [17].

3. Methodology

Our study can be divided into the following parts:

- Data collection
- Preprocess and prepare the dataset
- Employing clustering techniques

A flowchart of our overall analysis has been shown in Figure 1.

A. Data Collection

The microarray gene expression values used in this research were derived from the TCGA gene expression database. Six death groups, alive and one control group included in the dataset were studied. The groups studied are one month, three months, six months, one year, two years, three years later and alive. They include 20, 34, 76, 99, 144, 73, and 147 patients respectively. All six groups comprised of the dead, alive, and the control group had the same number of 17814 genes for the analysis. All groups had ID, death time an average of gene expressions, and gene ID.

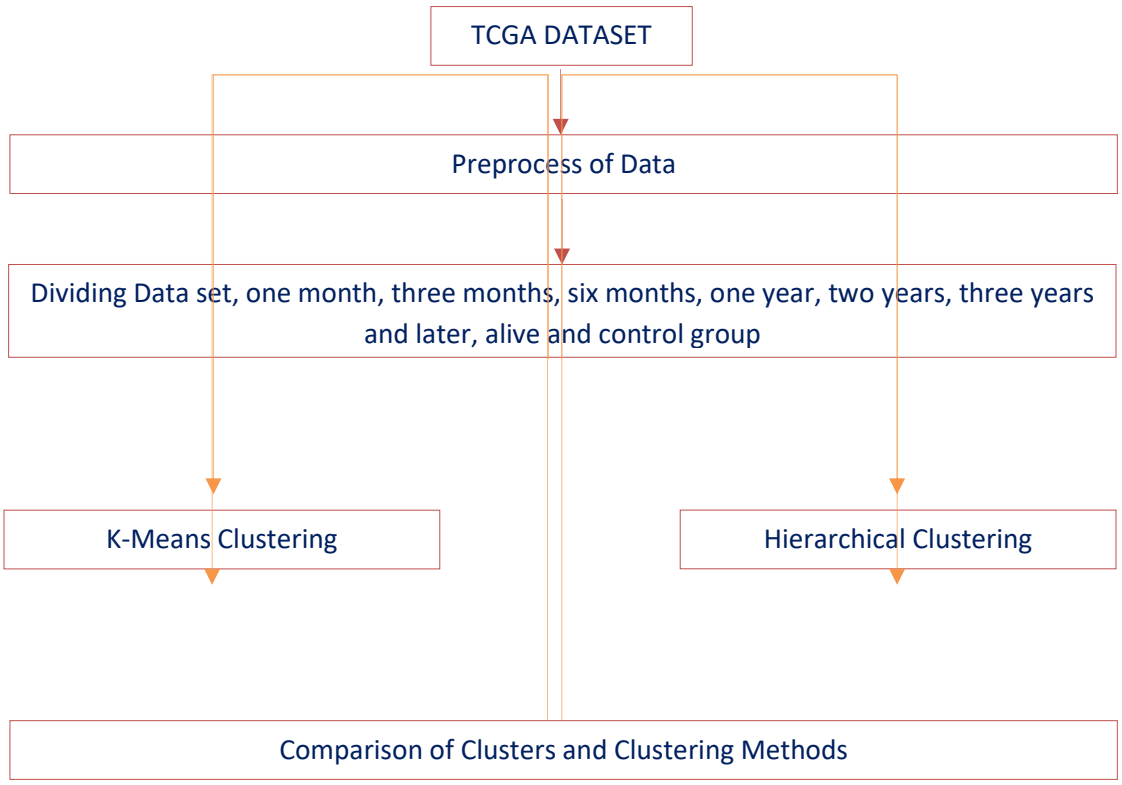


FIGURE 1. Flow-chart of our overall study.

B. Preprocess and Preparing Dataset

In the initial stage, patients whose data was missing in the clinical dataset were removed. Dead people's and alive patients' information were taken from the data set and dead people were classified into six different groups which were one month, three months, six months, one year, two years, and three years later depending on the duration of death time. Finally, other groups comprising of deceased, alive, and control groups were obtained.

4. Results and Discussion

A. Optimal Number of Clusters Calculation

To evaluate the optimal number of clusters for the clustering techniques, the Elbow method was used for K-means techniques while dendrogram was used for hierarchical techniques.

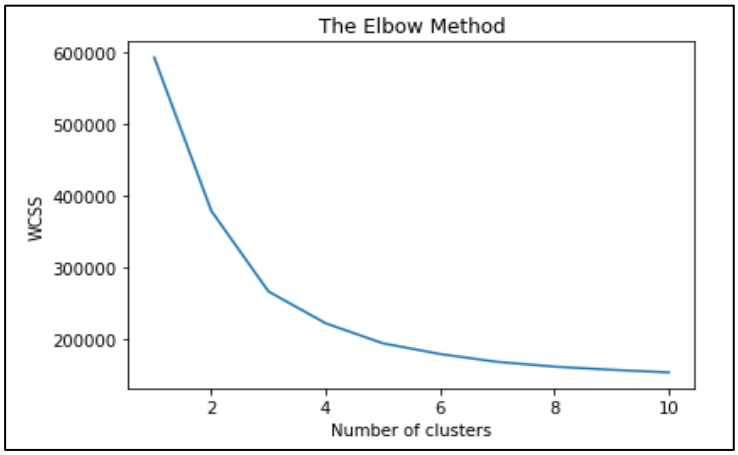
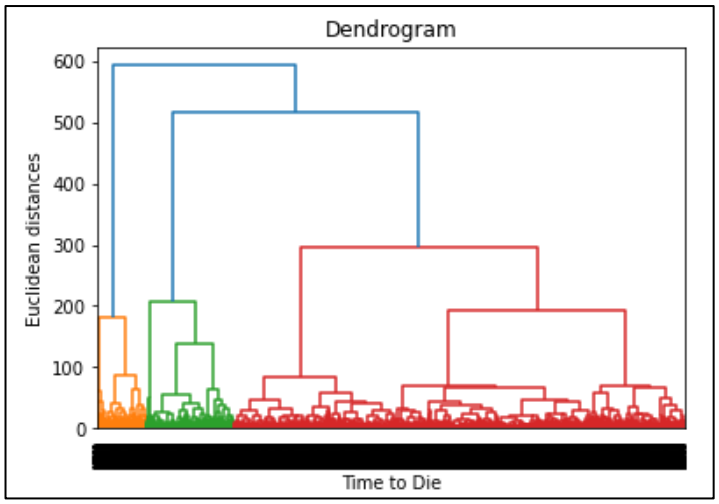


FIGURE 2. Elbow Method Curve for K-Means technique (for one month dataset).



3 clusters

FIGURE 3. Dendrogram for hierarchical clustering technique (for one month dataset).

B. *Dataset Evaluation*

The given graph in Fig. 4 shows a two-dimensional representation of the dataset which clustered the whole dataset that has 17814 genes separated into three groups using K-means clustering methods. These groups are 0th, 1st, and 2nd clusters. 69.32% of the genes in the control group belong to the 2nd cluster, 12.61% to the 1st cluster, and 18.07% to the 0th cluster (Table 1).

TABLE 1. Distribution Genes in the Clusters using K-Means Clustering Method.

Cluster	One Month	Three Months	Six Months	One Year	Two Years	Three Years	Alive	Control
0-Cluster	3578	12057	3420	11921	2361	11789	12088	3219
1-Cluster	11941	3445	2338	2307	11898	2240	3379	2246
2-Cluster	2295	2312	12056	3586	3555	3785	2347	12349

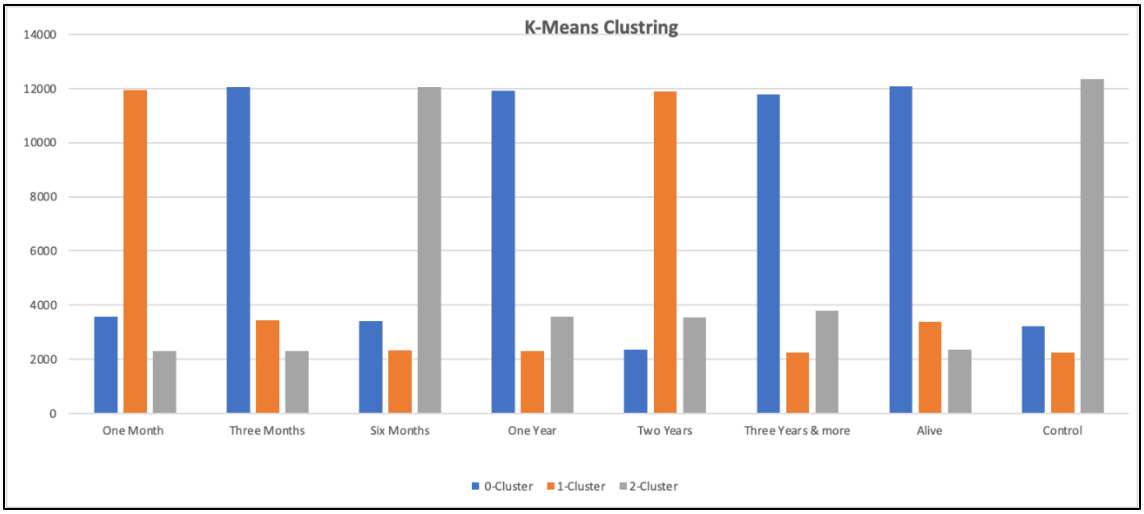


FIGURE 4. Clustering Dataset using K-Means Clustering Method.

From this graph, it can be observed that the second cluster contains more than half of the genes in healthy people. It has been determined that the cluster distribution of a healthy patient and a patient who died six months after being diagnosed with brain cancer is similar. Approximately 67% of the genes of healthy people and six months groups at the second cluster.

C. *Model Evaluation*

The developed gene expression clustering models were evaluated based on certain performance metrics such as classification based on min-max and average value. All these metrics were evaluated to ascertain the performance of the gene clustering model.

D. Min-Max Gene Expression Outcome

The clustering of gene expression data of brain cancer patients was evaluated in two-phase. The first phase of the clustering process was done by classifying each gene expression of each patient into three main outcomes: low, normal, and high based on their equivalent biosample ID and the aggregated minimum and maximum values of each type of gene as presented in Tables 2.

TABLE 2. Classification based on Min-Max Value.

Input	Rule	Outcome
Biosample repository ID	Value of biosample repository ID>Max	Low
Biosample repository ID	Value of biosample repository ID >Min, Value of biosample repository ID <Max	Normal
Biosample repository ID	Value of biosample repository ID >Max	High

E. Gene Expression Classes

The dataset was further divided into seven different groups based on the time spent by the patients after diagnosis, this was achieved by taking their arithmetic means. The average value of each gene was calculated by taking the average of the expression data, the outcome was classified into different groups. Any average value that is 0, belongs to class 0. The average value of 1, belongs to class 1. The average value of 2, belongs to class 2. The average value of -1, belongs to class -1. Lastly, an average value of -2, and belongs to class 2. In a case of greater than 2 average value, that is an extremely high class, and an average value greater than -2 is extremely low.

F. Gene Expression Classification

One of the objectives of this work is to classify the gene expression of brain cancer patients into three main (0, 1, 2) clusters of seven groups either 0, 1, -1, 2, -2, extremely low, extremely high. Fi 5 presents the percentage graphical representation of all the time spent by patients after being diagnosed with brain cancer in each class. Group 0 has the highest percentage in One month after the diagnosis dataset. Group -2 has the lowest percentage.

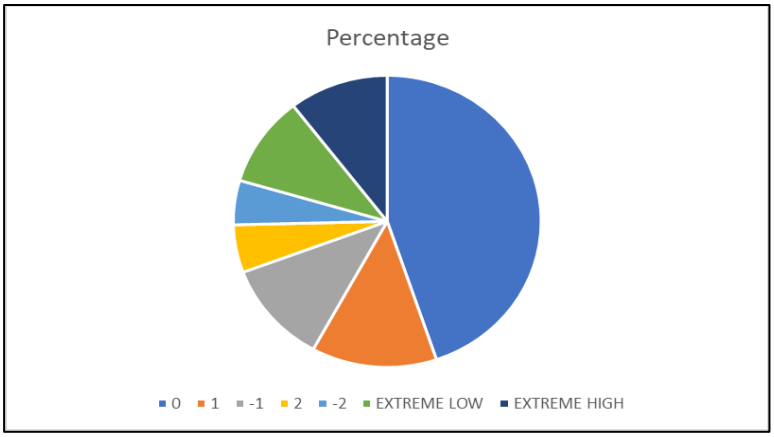


FIGURE 5. Gene Expression Classification for Patient Diagnosed with Brain Cancer One Month after Diagnosis.

G. Cluster Validity

This involves evaluating the clustering analysis results in a quantitative method. Clustering validity metrics are usually employed to calculate the optimal cluster numbers and it is usually dependent on clustering techniques employed. The idea behind cluster validity is to discover changing and non-changing clusters also known as compact and well-separated clusters. Compactness is employed to measure data variation present in a particular cluster, while the separation denotes segregation of clusters from one another. To achieve cluster validity, validity measures use sample mean of each subset, while others use all the points present in each subset in their computation.

H. K-Means Cluster Changing Rate

K-means clustering technique computes the centroids and iterates till optimal centroid is evaluated. The rate of change of cluster in the k-means technique is faster due to the quick identification of k number centroids that allocates every data point to the nearest cluster. The cluster comparison to control for clusters 0, 1, and 2 is 78.17, 8.77, and 13.06 respectively. This was evaluated by taking the percentage of a total number of genes in control for each cluster: 13925, 1563, 2326 respectively divided by a total number of genes.
Table 3 illustrates the cluster change for patient death duration after one month, each of the individual clusters has relatively the same cluster change pattern. For one month dataset using k-means clustering techniques, the cluster change rate is faster. The total number of genes present in the one-month dataset for cluster 0 is 3578, cluster 1 is 11941, and cluster 2 is 2295.

TABLE 3. One Month Death Duration Cluster Changing Rate for K-Means Technique.

Death Duration	**Method of Clustering**	**Clusters**	**Total Genes**	**Cluster Changing**
One Month	K-Means	0	3578	Yes
		0		No
		1	11941	Yes
		1		No
		2	2295	Yes
		2		No

I. Hierarchical Cluster Changing Rate

The cluster comparison to control for clusters 0, 1, and 2 are 78.17, 8.77, and 13.06 respectively. This was evaluated by taking the percentage of a total number of genes in control for each cluster: 13925, 1563, 2326. Presented in Table 4 is the cluster change for patient death duration after one month, each of the individual clusters has relatively the same cluster change pattern. For one month dataset using hierarchical clustering techniques, the cluster change rate is less fast compared with the K-means clustering technique. The total number of genes present in the one-month dataset for cluster 0 is 13679, cluster 1 is 1490, and cluster 2 is 2645.

The bar graph in Fig. 5 shows which group changed cluster depending on the control group. Compared to the genes in the control group and one month, three months, one year, two years, three years later, and alive groups have changed clusters from 72% to 79%. In the six months, just 47% of the genes had changed the cluster. The alive group has the highest rate at 78.89% in gene cluster replacement rates. The six-month group has the lowest rate, at 47.32%.

TABLE 4. One Month Death Duration Cluster Changing Rate for Hierarchical Technique.

Death Duration	Method of Clustering	Clusters	Total Genes	Cluster Changing
One Month	Hierarchical	0	13679	Yes
		0		No
		1	1490	Yes
		1		No
		2	2645	Yes
		2		No

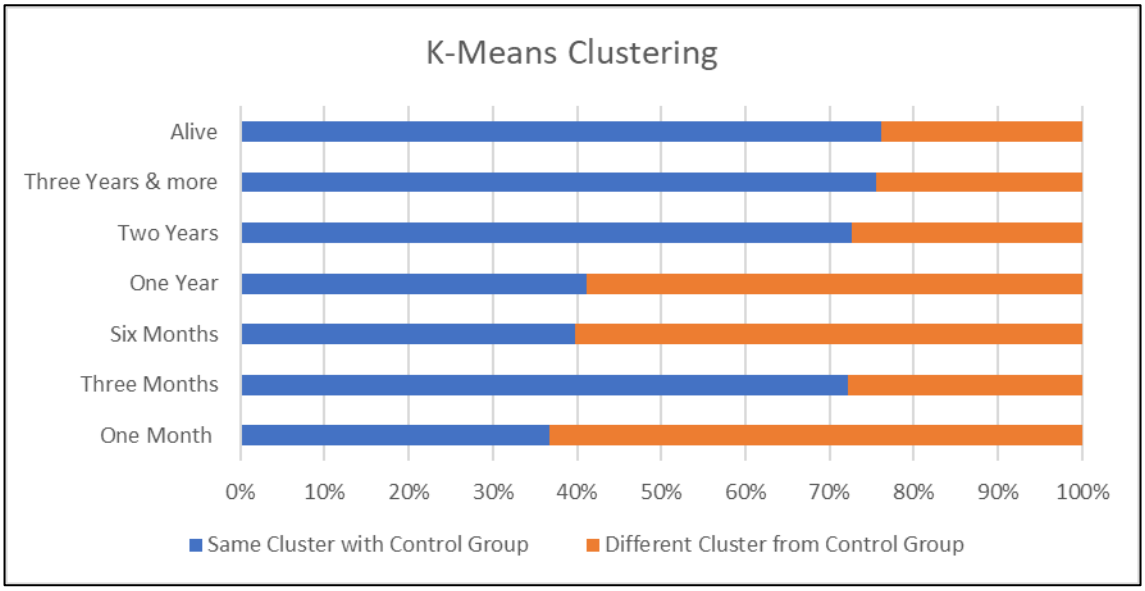


FIGURE 5. Shifting cluster between the control group and diseased groups.

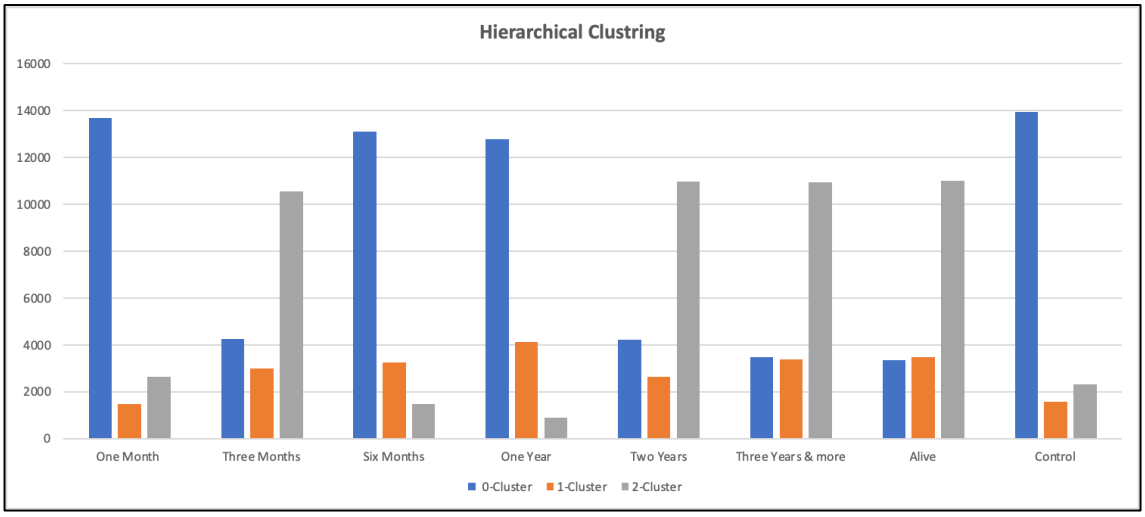


FIGURE 6. Clustering dataset using by Hierarchical Clustering Method.

The graph shows the distribution of genes using the Hierarchical Clustering method (Figure 6). The distribution of people in the control group using the Hierarchical Clustering method is in the table. 13925 of the genes are in the 0th cluster, 1563 are in the 1st cluster, and 2326 are in the 2nd cluster. In a healthy individual, the 0th cluster is seen as dominant. When the groups are observed, one month, six months, one year, and the control group have a similar distribution of genes, approximately 70% of genes in the 0th cluster. When the three months, two years, three years, and alive groups are examined, between 50% and 60% of the genes are in the 2nd cluster.

Table 5 shows the results of the hierarchical clustering method, including the number of gene clusters that have changed and how many of these are in the same cluster as their control group. According to the statistics in the table, the alive group had the greatest rate of people who changed clusters, at 76.19%. The one-month group had the smallest change of 36.80%.

TABLE 5. Distribution genes in the Clusters using Hierarchical Clustering Method.

Cluster	One Month	Three Months	Six Months	One Year	Two Years	Three Years & more	Alive	Control
0-Cluster	13679	4260	13105	12777	4220	3492	3346	13925
1-Cluster	1490	3001	3241	4121	2633	3376	3476	1563
2-Cluster	2645	10553	1468	916	10961	10946	10992	2326

6. Conclusion

Using clinical data and gene expression data from its database, seven distinct groups, and control groups were formed in this study. In the formed groups, three different clusters were obtained by using K means and Hierarchal clustering methods. In these two approaches, three clusters were also used to create gene expression in the control group. Most of the genes in the control group were in the second cluster. On the other hand, according to the data obtained using the K-means clustering approach, most of the genes in patients with the disease differed in the cluster distribution. In the result of the Hierarchical Clustering Method, most of the genes in the control group were in the 0th cluster and three groups were found to have a similar distribution to clusters in the control group. Except for these three groups, the gene distribution of the other groups differed. When K-means Clustering Method was compared with the Hierarchical Clustering Method, it could be found that the cluster change rate according to the control group was higher in the K-means Clustering Method. In the future, work can be done using different machine learning clustering methods. By using Clustering methods and extreme gene expressions, it can be revealed which genes are effective in brain cancer.

7. References

[1] World Health Organization (WHO). Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. WHO; 2020. [Accessed 16 June 2021]. who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

[2] Amyot, F., Arciniegas, D. B., Brazaitis, M. P., Curley, K. C., Diaz-Arrastia, R., Gandjbakhche, A., Herscovitch, P., Hinds, S. R., Manley, G. T., Pacifico, A., Razumovsky, A., Riley, J., Salzer, W., Shih, R., Smirniotopoulos, J. G., & Stocker, D. (2015). A Review of the Effectiveness of

Neuroimaging Modalities for the Detection of Traumatic Brain Injury. *Journal of Neurotrauma*, *32*(22), 1693–1721. https://doi.org/10.1089/neu.2013.3306

[3] Bray, F., Laversanne, M., Weiderpass, E., & Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, *127*(16), 3029–3030. https://doi.org/10.1002/cncr.33587

[4] de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: A comparative study. *BMC Bioinformatics*, *9*(1), 497. https://doi.org/10.1186/1471-2105-9-497

[5] Dirks, P. B. (2008). Brain Tumor Stem Cells: Bringing Order to the Chaos of Brain Cancer. *Journal of Clinical Oncology*, *26*(17), 2916–2924. https://doi.org/10.1200/JCO.2008.17.6792

[6] Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine Learning for Medical Imaging. *RadioGraphics*, *37*(2), 505–515. https://doi.org/10.1148/rg.2017160130

[7] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, *286*(5439), 531–537. https://doi.org/10.1126/science.286.5439.531

[8] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[9] Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Edition 1). STHDA.

[10] Mahaley, M. S., Mettlin, C., Natarajan, N., Laws, E. R., & Peace, B. B. (1989). National survey of patterns of care for brain-tumor patients. *Journal of Neurosurgery*, *71*(6), 826–836. https://doi.org/10.3171/jns.1989.71.6.0826

[11]Morris, Z., Whiteley, W. N., Longstreth, W. T., Weber, F., Lee, Y.-C., Tsushima, Y., Alphs, H., Ladd, S. C., Warlow, C., Wardlaw, J. M., & Al-Shahi Salman, R. (2009). Incidental findings on brain magnetic resonance imaging: Systematic review and meta-analysis. *BMJ*, *339*(aug17 1), b3016–b3016. https://doi.org/10.1136/bmj.b3016

[12] Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: An overview, II. *WIREs Data Mining and Knowledge Discovery*, *7*(6). https://doi.org/10.1002/widm.1219

[13] Nidheesh, N., Abdul Nazeer, K. A., & Ameer, P. M. (2018). *A Hierarchical Clustering Algorithm Based on Silhouette Index for Cancer Subtype Discovery from Omics Data* [Preprint]. Bioinformatics. https://doi.org/10.1101/309716

[14] Tan, P.-N., Steinbach, M., & Kumar, V. (2006a). *Introduction to data mining* (1st ed). Pearson Addison Wesley.

[15] Tan, P.-N., Steinbach, M., & Kumar, V. (2006b). *Introduction to data mining* (1st ed). Pearson Addison Wesley.

[16] Tan, P.-N., Steinbach, M., & Kumar, V. (2010a). *Introduction to data mining* (Pearson internat. ed., [Nachdr.]). Pearson Addison-Wesley.

[17] Tan, P.-N., Steinbach, M., & Kumar, V. (2010b). *Introduction to data mining* (Pearson internat. ed., [Nachdr.]). Pearson Addison-Wesley.

[18] Tandel, G. S., Biswas, M., Kakde, O. G., Tiwari, A., Suri, H. S., Turk, M., Laird, J., Asare, C., Ankrah, A. A., Khanna, N. N., Madhusudhan, B. K., Saba, L., & Suri, J. S. (2019). A Review on a Deep Learning Perspective in Brain Cancer Classification. *Cancers*, *11*(1), 111. https://doi.org/10.3390/cancers11010111

[19] VASANTHA, M., SUBBIAH BHARATHI, & SUBBIAH BHARATHI. (2010). Medical Image Feature, Extraction, Selection And Classification. *International Journal of Engineering Science and Technology*, *2*(6), 2071–2076.

Recommendation Engine on IPTV

Vedad Njuhović*, Samed Jukić*
*International Burch University,
vedad.njuhovic@stu.ibu.edu.ba
samed.jukic@ibu.edu.ba

Original research

Abstract: *In recent years IPTV (Internet Protocol Television) platforms are becoming one of the most popular entertainment multimedia services which are used to serve movies, tv-series and other video and audio attractive content using the Internet Protocol. VoD (Video on Demand) is the most popular multimedia IPTV service, which provides content without the need for the old traditional way of using video playback devices. Except that it is necessary to have high-quality VoD content, IPTV platforms must provide the best end-user experience. Moreover, it is imperative to provide new features to attract new customers and keep the existing ones. We confirmed the efficacy of this classifier thru simple trial and error. When we searched for movies that have sequels, our engine recommended those sequels. Since Cosine Similarity Classifier considers multiple factors, such as actor, genre, year, etc. Even if the movie does not have prequels or sequels this algorithm was able to provide us with movies that share other common characteristics.*

Keywords: IPTV, engine, unicast, broadcast, multicast, python, CSS, HTML, JS.

1. Introduction

In the next couple of chapters, this paper will explain what IPTV is, what is a Recommendation engine, and how a recommendation engine can be used to improve the end-user experience. In the first part, we will define general terms which need to be understood before proceeding. The second part will cover current solutions which are implemented on market for this purpose. The methodology will be covered in the third part, which includes technologies used for the implementation of this project. The last part will summarize the results obtained after the implementation of this system.
IPTV stands for "Internet Protocol Television". The "IP" in IPTV is the same as your IP or VoIP (Voice over IP) address. All this means that television programs are distributed using Internet protocols [1].
A recommendation engine is a data filtering tool that uses machine learning algorithms to recommend the most relevant items to specific users or customers. It works by looking for patterns in consumer behavior data, which can be collected implicitly or explicitly [2-5].

2. Literature Review

It is really difficult to enumerate all the different fields in which recommender systems are applied. For example, they are used in e-commerce [6], [7], for recommending music, movies, scientific papers, etc. GroupLens [8] was one of the first implementations of recommender systems. Unlike our hybrid technique, it uses the traditional collaborative filtering approach to calculate new recommendations. Later, the same research group developed MovieLens, a movie recommender system that is also based on traditional collaborative filtering. We proposed an approach based on Cosine Similarity Classifier to increase recommendation accuracy. Based on this approach, we developed a movie recommendation system that combines both content-based and collaborative information about movies. This system takes into account all separate information about a movie such as actors, year, genre, etc, and by searching most similar items based on multiple factors it delivers the most similar movies.

A more interactive movie recommender system, named MovieGEN is presented in [20]. This is a hybrid recommendation system, which uses machine learning and cluster analysis to calculate recommendations. However, this system uses the personal information of users to predict their movie preferences using well-trained support vector machine (SVM) models. Then, based on SVM predictions, it selects movies from the dataset, clusters them, and generates questions for users. Finally, it uses the information collected through the answers to refine user recommendation lists. The system recommended by Conti's IPTV [5] is a recommended social system based on the profiles of the social network and the analysis of the activity of personal users according to similar activities. Social recommendations. In addition, we analyzed the information of the live program of Electronic Program Guide (EPG) and we made the time division for recommendations to distinguish users. Jinni [8] is a semantic search engine for movies and television content and a recommendation of them. Another engine analyses only the information of the standard film (as titles, actors, lists, director, scenarios, etc.) approach such as this one, enables significant content searches and allows quick adaptations to the user's benefit. The authors are based on the implicit feedback of the user who divides the user profile into multiple subfiles called Microfofiles based on the user, each based on multiple subfiles representing the user's technology introduced. Two different context Recognition technologies for film recommendations are shown in [7]:

- Greater performance of conventional C.
- Approaches that use electrical contextual time factors.
- Implementation of machine learning.

The authors propose the use of temporary information to improve the recommended quality. For example, [13] proposes a technique to model the temporary dynamics of client preferences by separating the transient factors from the last. The recommendation service of Recommendation, based on the interesting location, which takes a Temporary context, is [15], but [16] is feedback from implicit users and information on the user's purchase time and the start-up of the article to achieve the accuracy of the recommendation.

3. Methodology

A. Acquisition

To prepare the recommendation engine, data should be collected, cleaned, and prepared for the engine needs. In this project, we have collected data from 6 different sources, cleaning of data was performed in 5 pre-processing steps which are going to be explained in detail below.
Pre-processing 1:
The first dataset was retrieved from Kaggle and is called "IMDB 5000 Movie Dataset" with a CSV file called "movie_metadata.csv".
Movie_metadata.csv consists of movie information with columns such as director name, main actors, movie name, genre, number of reviews, number of likes, duration, etc. From all the columns we will keep only the most important columns such as 'director_name',' movie_title', 'genres' etc.

Since in each column there are a lot of NaN values which are replaced with 'unknown'.
In this dataset genres are written with pipe '|', so it should be replaced with just blank/space instead, for the better processing of the data. Once all NaN values are replaced it is time to transform all movie titles to lower case. After completing this step, we realized that each movie name consists of \xa0 (single character) which should be removed to have correct names. Using the lambda function (iterating each movie title) and using the syntax [:-1] we are excluding the last part of the movie titles. Finally, the last step is saving cleaned and prepared data to the common CSV file.

B. Pre-processing 2

Since the first movie dataset included data till 2016, this dataset includes data from 2017. So using this dataset we will take data of the 2017 year and merge it with the first dataset. This dataset has separated actors and directors in another dataset called 'credits.csv', this data should be merged with the main dataset by id. Genres in this dataset are like an array of objects but in form of a string, they should be converted like in the first dataset, like genres with space-separated. For converting this, we have used AST (abstract syntax trees) and 'literal_eval' which constructs an object from string. Using a custom function for iterating through the objects in the array we achieved the same result for genres like for the first dataset. The same process is repeated for extracting actor names and also for the directors.
After this, we need to take only the most important columns, like for the first dataset with dropping NaN values and renaming all columns to match the first dataset.
The most important step is the creation of a new column called 'comb' which consists of director names, actor names, and movie genres. The last but not least step is combining the first and second movie datasets into the same file.

C. Pre-processing 3

The first and second data set was taken from Kaggle and includes movies till 2018, so we decided to take movies from Wikipedia by extracting a list of films in 2018. After taking all the

movies from Wikipedia, I realized that there is no genres column. Using the tmdb3api for each movie by its name retrieve its id which is used for the new request for getting genres.
Since the cast and crew column included all information including directors and actors we needed some custom functions with basic splitting for retrieving its names. Last but not least step is renaming all columns to match the previous dataset. The last step is the creation of a comb column with merging all previous data with 2018 data.

D. Pre-processing 4

In pre-processing 4, we have performed the same steps as in pre-processing 3 for retrieving movies of 2019 and 2020.

E. Pre-processing 5

In pre-processing 5, we have performed the same steps as in pre-processing 3 for retrieving movies of 2021.

F. Similarity matrix

After the prepared data, we are ready to move forward with the creation of a recommendation engine.
Our dataset contains as we already explained info such as movie name, genre, casts, director, etc. The last field in the dataset contains a field that has all this data combined, as plain text.
Let us consider that our dataset has only 10 movies so we simplify the explanation. Since we have 10 movies, we will generate a matrix of size 10x10, where for example element [3], [5] quantifies how much movie 3 is similar to movie 5. These quantities are calculated by cosine similarity, which is explained earlier in this chapter two plain texts are compared, each movie is compared with each one, and then the algorithm calculates similarity as a number between 0 and 1. These calculations are only performed once, and this matrix is saved.

Now consider that the user is searching for movie number 7 in our list (that is movie from our 10-movie dataset which we mentioned above). Our program will take all movies from row number 7 because that row contains similarity of movie 7 with all other movies in our database. If the numeric value in the field is higher that means the movie is more similar to the one user searched originally. If we want to get the top 5 most similar movies, we will request the top 6 movies, and ignore or remove the one with similarity one, or the most similar movie, because this is the value where movie number 7 was compared to itself, or in our case that would be [7].

4. Results

In the figure below, the homepage of our application is presented.

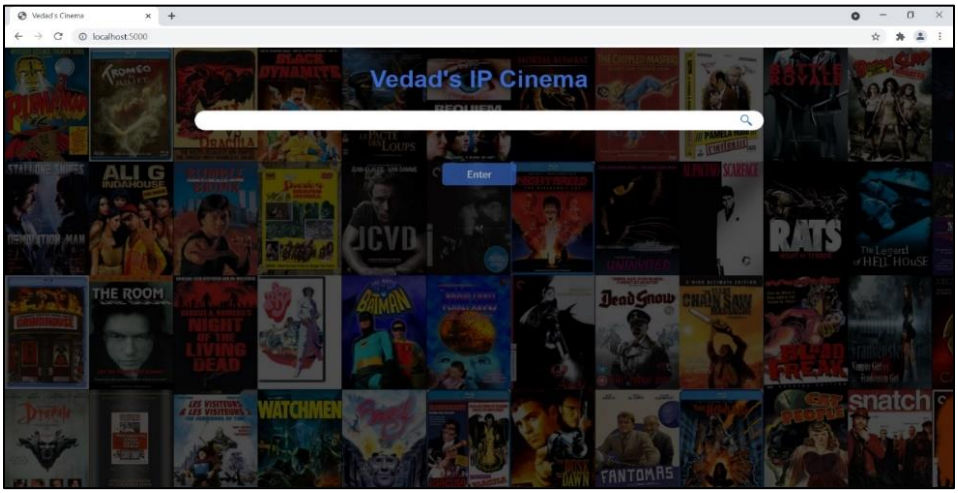


FIGURE 1. Homepage.

The user has can browse thru movies. After entering three letters recommended movies start to pop up. Recommendations or autocomplete are obtained by calling the function "get_suggestions()", which gets data from AJAX requests. After the user selects a movie, the screen from figure 2 below appears.

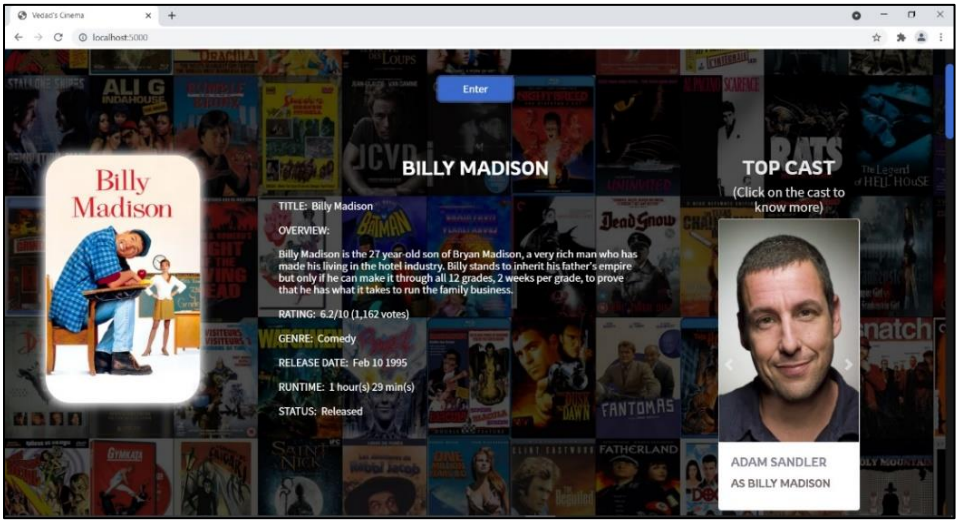


FIGURE 2. Movie page.

If we scroll down, we may see user reviews, which are obtained from IMDB. Figure 3 contains reviews and is presented below.

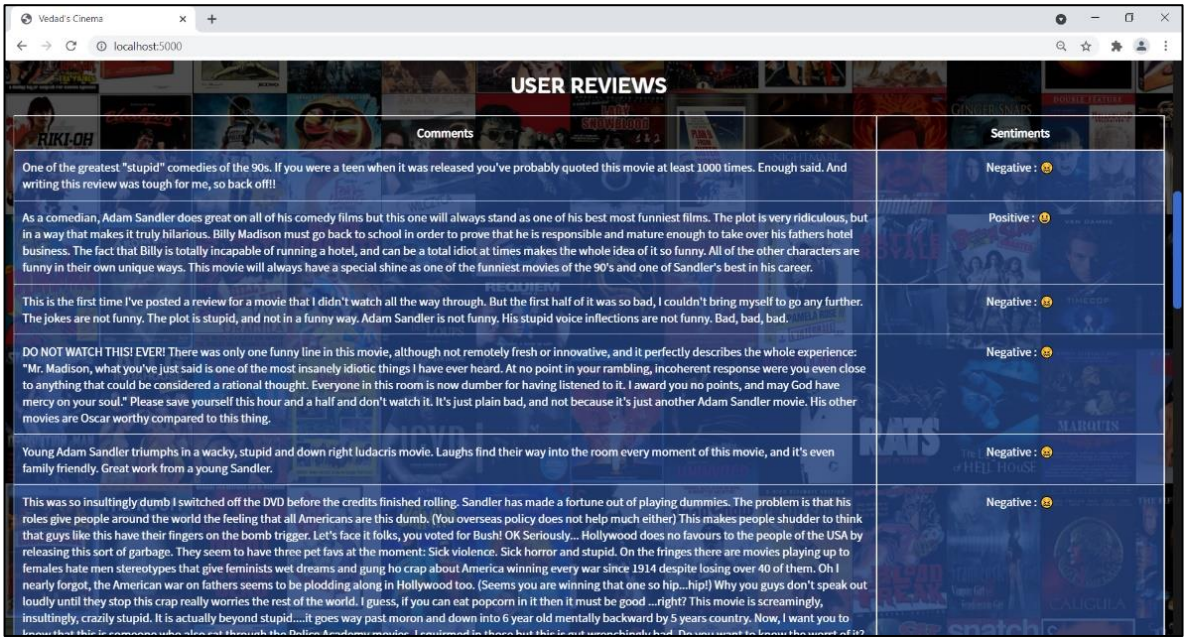


FIGURE 3. Representation of user reviews.

At the bottom of the page, we may find suggested movies, which are delivered by API based on genre, actors, and similar factors.

A. Accuracy

Since the recommendation engine needs to be personalized toward users' requirements, we surveyed a small group of people and asked them which movies they expect to be recommended if they searched for certain movies.
Survey returned the following results:

- System recommended 1 out of 4 expected movies 2 times
- System recommended 2 out of 4 expected movies 5 times
- System recommended 3 out of 4 expected movies 38 times
- System recommended 4 out of 4 expected movies 5 times

Graphical representation of results is presented in Figure 4

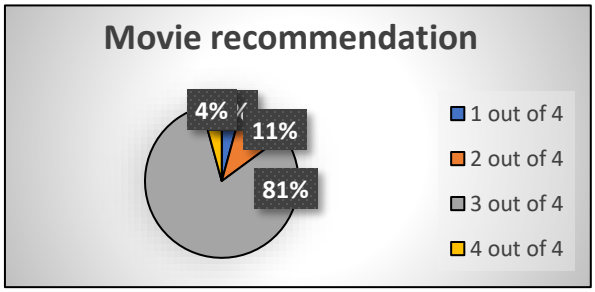


FIGURE 4. Pie chart representing survey results

5. Conclusion

IPTV as a concept first time appeared in the 1990s, and it presented a way of delivering video content over networks [1]. The factors such as reliable package delivery protocol and video compression algorithms allowed this technology to be developed in commercial installations. Primary underlaying protocols which are used in stand-based IPTV are [3]: Unicast web-based live and VoD streaming and Web-based multicast. A set of hardware equipment and software which are interconnected via a network is called an IPTV system [5]. Video on demand servers pr storage for media such as tv shows, video clips, movies, and similar. Their job is to provide secure and perfect access to content that is stored [8]. Session level protocol which is defined by the IETF to provide transportation functions over the network that facilities the delivery of data such as video or audio in real-time using either unicast or multicast technology is called Real-Time Protocol. Technologies that we used for the development of this project are HTML, CSS, JS, AJAX, and Python. Cosine Similarity presents a measurement that quantifies the similarity between at least two vectors**.**

Considering that we spent most of our time preparing the data for processing, it is clear how important this is for the modeling itself. In comparison with the already made recommendation engines, we have shown that in a much simpler way it is possible to make a good enough recommendation engine, which proved to be of good quality taking into account the statistics of the respondents.

The main reason why I chose this topic and embarked on the adventure of making a recommendation engine is the current trend of watching Netflix and the marathon search for a movie we would like and would love to watch. Taking into account the result of the project, I managed to reach the goal I had set.

6. References

[1] International Telecommunication Union ITU, «Agentia National Pentru Reglementare», January 1998. [En línea]. Available: http://www.anrceti.md/files/filefield/Recomandarea%20ITU%20H.323_0.pdf.

[2] Committed to connecting the world, «Telecommunication Standardization Sector», ITU, [En línea]. Available: http://www.itu.int/en/ITU-T/Pages/default.aspx.

[3] I. K. Park, O. Seung Hun, Y. S. Kwon and H. Young Song, «An implementation of user-participated interactive IPTV service system», IEEE International Symposium on consumer electronics, Perth, 2008.

[4] K. Kerpez, D. Waring, G. Lapiotis, J. Lyles and R. Vaidyanathan, «IPTV service assurance», IEEE Communications Magazine, vol. 44, n°9, pp. 166-172, 2006.

[5] G. Myoung, L. Chae, S. Lee, W. Seop Rhee and J. Kyun Choi, «Functional architecture for NGN-based personalized IPTV services», IEEE Transactions on Broadcasting, vol. 55, n°2, pp. 329-342, 2009.

[6] Arnold, D., Bond, A., Chilvers, M., & Taylor, R. (1996, June). Hector: Distributed objects in python. In *Proceedings of the Fourth International Python Conference, Livermore CA*.

[7] P.S. Pacheco, Parallel Programming, Morgan Kaufmann Publishers, 1997.

[8] López Sarmiento, D. A., Villanueva Ocampo, B. F., & Rivas Trujillo, E. (2013). Iptv: Next-generation network technologies and protocols. *TECCIENCIA*, *7*(14), 39–48. https://doi.org/10.18180/tecciencia.2013.14.5

[9] Mirri, S., Peroni, S., Salomoni, P., Vitali, F., & Rubano, V. (2017). Towards accessible graphs in HTML-based scientific articles. *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, 1067–1072. https://doi.org/10.1109/CCNC.2017.7983287

[10] Zeadally, S., Moustafa, H., & Siddiqui, F. (2011). Internet Protocol Television (IPTV): Architecture, Trends, and Challenges. *IEEE Systems Journal*, *5*(4), 518–527. https://doi.org/10.1109/JSYST.2011.2165601

Depression and Anxiety Analysis and Prediction using Decision Tree and Logistic Regression

Mersiha Ćeranić*, Samed Jukić*
*International Burch University
mersiha.ceranic@stu.ibu.edu.ba
samed.jukic@ibu.edu.ba

Original research

Abstract: *COVID-19 pandemic brought many changes in people's lifestyles. Some of those changes hurt people's mental health in different age groups. This research is done to investigate which factors contributed most to the occurrence of depressive and anxiety symptoms during COVID-19 lockdown, and what type of people in terms of age, sex, level of education, place of living, was the most exposed to the appearance of mental health disorders. 1115 people (18-85 years old) from Poland joined the research process. They fulfilled online questionnaires which were used as a basis for further research of lockdown impact on mental health. Responses are evaluated by using ML tools predicting the group of participants with signs of depression and anxiety, based on their answers to the questionnaires, and the attributes of the participants. Based on the results given by the studies, the youngest population (age 18-29), which participated in the surveys, experienced more intense depression and anxiety symptoms than participants from other age groups.*

Keywords: Anxiety, covid-19, decision tree, depression, logistic regression, tableau, Weka.

1. Introduction

During the COVID-19 lockdown, many people were facing mental health issues, where the global pandemic, caused by the spread of the COVID-19 virus, has changed statistics for mental illnesses, depression, and anxiety. The percentage of the population who suffer from mental illnesses increased significantly during the COVID-19. There were different factors, such as difficulties in family relationships, lack of live communication within the community, insecurity as the virus spread, etc., causing people to feel depressive and anxiety symptoms. Various studies have shown that people at a younger age were more affected by having depression and anxiety symptoms than the older ones throughout the lockdown. Relationship difficulties and bad communication at home are determined to be the most common factors of having anxiety and depression symptoms among all participants throughout the pandemic, regardless of age. Besides recently mentioned, frequent factors of mental health issues between the youngest throughout the pandemic were impediments related to constraints outside [1].

This research is conducted to determine which factors have the greatest impact on the occurrence of depression and anxiety symptoms among the population, considering their age, gender, financial situation, etc. Besides that, we wanted to investigate the effects of the COVID-19 pandemic on the mental health of people aged 18-85, and the possible long-term consequences of the lockdown on it. This will help prevent the occurrence of further bad outcomes for people's mental health.

The result of this research process will be the implementation of Machine Learning models, which will classify the people who participated in the research process into two classes: those who have symptoms of depression and anxiety, and those who have not. Classification is based on the attributes/characteristics and the answers provided by the people who participated in these studies.

The following chapter, after “Introduction”, is “Literature review”, where we will make an overview of similar studies. Chapter “Methods” will represent the process of chain decisions and research that will shape the final solution. Surveys that are used for gathering participants’ characteristics, will be described here. Within this chapter, we will also give a brief overview of data used in this work, but also the steps that data needs to go through before creating a Machine Learning model. This chapter will give a brief overview of the Machine Learning classification algorithms that were used for the implementation of Machine Learning models. Besides that, this chapter is providing information about the tools used during the work. Chapter “Results” will present how accurate are Decision Tree and Logistic Regression ML models are in detecting which records, based on their features and answers to the online questionnaires, might have symptoms of depression and anxiety, and which records might not have. Besides that, this chapter also explains in detail the performance evaluation of implemented Machine Learning models, by using accuracy parameters. The chapter “Discussion” will be presented a comparison between our studies and similar ones. The “Conclusion” chapter will conclude our studies.

2. Literature Review

Mental wellbeing during coronavirus disease (COVID-19) has negatively impacted many people around the globe. Reports suggest that almost four times as many people stated clinically important signs of depression or anxiety during January 2021 than in January through June 2019 [2]. Prevention and control measures such as social distancing, closing down businesses and distance learning created negative conditions which lead to negative consequences for persons' mental health. The COVID-19 depression rate is different among different groups of people. For example, youths will more often notify depression or anxiety signals than older age groups, and people with lower income are more likely to declare that stress related to the COVID-19 period had a major negative influence on their mental wellbeing.

Alarming levels of unfolding and severity lead the World Health Organization (WHO) to declare the coronavirus (COVID-19) natural event a worldwide crisis on March 11, 2020. This caused several countries around the world to start out implementing countermeasures within the style of limiting people's movement and interactions like the closure of schools, kindergartens, and nurseries. Closing down schools and businesses, as well as isolation from friends and relatives, brought many concerns about the mental state of younger ones. Constraints caused by a pandemic impacted the economy negatively, which brought many difficulties in maintaining mental balance. Parental job loss and financial insecurity also contributed to psychological distress becoming a rising trend among youth.

The overall aim of these studies is to identify and explore the symptoms of anxiety and depression before and during the COVID-19 period. In this respect, the research objectives that arise from the review of the literature is to do more research on which factors contributed most to the occurrence of depressive and anxiety symptoms during COVID-19 lockdown, and what type of people in terms of age, sex, level of education, place of living, was the most exposed to the appearance of mental health disorders.

Women, especially younger ones are more exposed to perceiving anxiety and depression during elevated stress periods [3]. Job loss, financial insecurity, fear of contracting the virus, and consequences of lockdown restrictions were high contributors to the poor mental state between men and women, regardless of age [4]. Many people across the world are affected by mental issues. The types of mental illness are placed in the world for a long time. Depression and anxiety are the most common types of mental disorders [5].

Anxiety can include concerns about problems such as money, health, and/or family problems. People who suffer from anxiety are extremely worried about the mentioned or some other things, even when there is no need for them. They are very worried about how they will get through the day, think negatively, underestimate themselves, and think that all things will go wrong. Studies gave the results that in 2017, American patients, mostly affected by anxiety, were aged 30 – 44. The following group was 18 – 29, which presents 22.3%, and group 45 – 59 presents 20.6%. People aged 60+ were the least affected in 2017.

Illnesses like depression are feelings that cause an enduring feeling of unhappiness and/or loss of interest. Depression is additionally referred to as affective disorder or emotional. It greatly affects people's behavior, thinking, and feeling. Those that suffer from depression might have issues with daily activities, and additionally to the flow, they often feel that living isn't worthy. Quite 264 million people worldwide suffer from depression. Someone stricken by depression will suffer severely and perform poorly at work, college and family. Worst of all, a mental state like depression can result in suicide. About 800,000 people, typically aged 15 -19, die thanks to suicide yearly. Between 76% and 85% of individuals in poor income, counties have no treatment for their disorder [6].

Statistics for sickness, depression, and anxiety are being modified thanks to the COVID-19, wherever for the studies were participated 398,771 patients. The share of the population who suffer from mental illnesses was increased during the COVID-19, where the report for the sickness gave the results: 28% for depression, ~27% for anxiety, ~24% for post-traumatic stress symptoms, ~37% for stress, 50% for psychological stress and ~28% for insomnia issues.

A. Related work

In an article by Marco Delmastro & Giorgia Zamariola, we can find most likely the first study of the impact of COVID-19 on mental well-being on a random and representative sample of the population of Italy. As we all know COVID-19 had a tremendous influence on social, personal and many other aspects of our lives and Italy was one of the first countries in Europe that experienced large-scale issues in the health sector due to COVID. This research included more than five thousand individuals who shared their gender, age, location, living situation, and socio-

economic status were also considered. Italy was the first country in Europe that reported the first death related to COVID-19 which took place on February 21st, 2020. Following this Italy was also the first country in Europe which initiated full-state lockdown measures due to the rapid spreading of the infection and the number of deaths. The lockdown period included travel restrictions, closure of schools, and nonessential industries. Requests to stay at home and avoid live meetings with people to inhibit the spreading had the potential to impact people's mental state in an extremely negative way, fear of getting infected, worry about death, and anxiety about uncertainties all contributed to worsening of mental state. People's exposure to constant COVID-related news also produced negative effects since much of that information was misleading or harmful. This data was collected using online questionnaires and social media. The study focused primarily on adults and young adults who completed self-administered questionnaires. The data collection was conducted using a mix of CATI and CAWI methods to limit the risks of self-selection and sample distortion. All the research happened based on the Declaration of Helsinki, it was also approved by the ethics commission. Data that was used for research was provided by the national department of civil protection. COVID-19 crisis was leading to the sharp rise of the number of people with poor economic status. The study showed that mental health is just as important as the physical when it comes to one's well-being [7].

Article by Ayesha Kamran Ulhaq, Amira Khattak, Noreen Jamil et al. (2020), describes in which way Data Analytics and its features can contribute to handling mental disorders in the right way. Worldwide, many people are suffering from mental issues daily. Adequate therapy and prevention would make their lives much easier. This is where Big Data comes into place. Researchers are working hard to find appropriate solutions that will predict future data, so the outcome would serve its purpose – prevent the increase of mental health issues among the population and predict mental illness. Through this article, it is displayed how Data Science is helping in predicting mental illness by utilizing Artificial Intelligence and Big Data, and also how mental disorder occurrence can be predicted by using Smart Devices. By this research, it is presented how social media is playing a very important role in predicting mental disorders. They were using Twitter data to realize these studies. Besides, the authors also have shown what challenges we can face when working with Big Data. By using Machine Learning algorithms for predicting future states/data to know how mental states can be improved or worsened, many lives can be saved in the future. Properly dealing with mental health issues will reduce the death rate and save many people from harming themselves, or at worst committing suicide [8].

In the article by Robert Stewart and Katrina Davis (2016), great interest in researching 'Big Data' resources are found in healthcare. However, to date, the applications of 'Big Data in the mental states have been left significantly restricted. When it comes to Big Data, the big challenge comes with the size of data sets, the speed of data collection, and the diversity of data. Numerous examples can be used for health studies, including the ones that are taken from huge collections of biological samples, research with a lot of complexity, social media, and so on. Clinical notes are also potential resources for Big Data when transferring data from paper to electronic format. Over the years, the "routine" has become the use of mental health data, where one can speak of asylum records in the mid-late twentieth century. However, with digitization and modernization, most data today is accumulated in electronic format. A large range of Bid resources then emerges as "mental health research platforms". Inevitably, the features of the resources will raise questions about the availability of such data. Some databases ret full hospital information from the electronic health record from the hospital. Some of the are populated from particular info which is used from the healthy grieve service department for the research procedure. Also, some data vas unmodified administrative data, and some databases rely on patient statements and reports [9].

3. Methods

A. Questionnaires

Depression and anxiety symptoms existence was assessed by the subsequent questionnaires, respectively: Patient Health Questionnaire-9, Generalized Anxiety Disorder-7, Scale of Perceived Health and Life Risk of COVID-19, Social Support Scale, Scale of Pandemic-Related Difficulties. The following part represents the psychology analysis of questionnaires.

- The Patient Health Questionnaire-9

Consisting of 9 fundamental things and 1 extra unit, it is used as an assessment instrument of chance for depression occurrence. Participants in the study gave answers from 0 - not at all to 3 - nearly every day.

- The Generalized Anxiety Disorder-7

Consisting of 7 things, it is used as an assessment instrument of chance for any disturbance occurrence, with the focus on concerning the frequency of signals throughout the last 14 days. People who participated in the studies answered from 0 - no to 3 - almost every day.

- Scale of Perceived Health and Life Risk of COVID-19

This survey was based on the research of the following possible threats: 1) COVID-19 contagion; 2) health problems that can be caused by a virus contagion; 3) life hazard as the outcome of the contagion. Every research field mentioned above had focus on the following things: the first one relates to the participants personally and the sec one to their loved ones. They answered from 1 - very low to 5 - very high.

- Scale of Epidemic-Related Difficulties

In this survey, people who joined the questionnaires were focused to give the answers to which pandemic-related difficulty they are dealing with at the moment. They answered from 1 - not an issue for me to 4 - it is an issue for me at all.

- Social Support Scale

In this survey, people who participated in the studies gave the answers if they are getting any of the following three kinds of social support: mental, physical, or communication and gentle/kind support. They gave answers from 1 - not getting any kind of support to 5 - yes.

B. Participants' Features Overview

Depression and anxiety analysis was done among 1115 participants, aged 18-85. Studies enclosed 563 females (50.5%) and 552 males (49.5%).

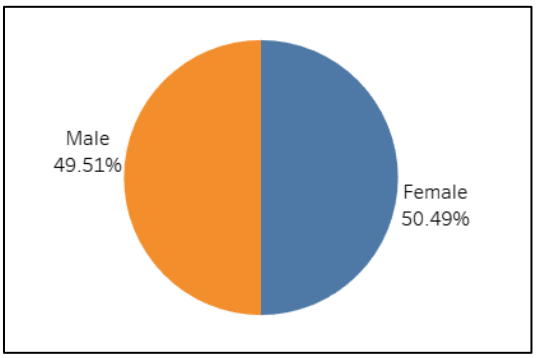


FIGURE 1. Percentage of participants by sex.

Studies brought together participants from four age groups: emerging adulthood (18-29 y.o.), established adulthood (30-44 y.o.), middle adulthood (45-59 y.o.), late adulthood (60-85 y.o.).

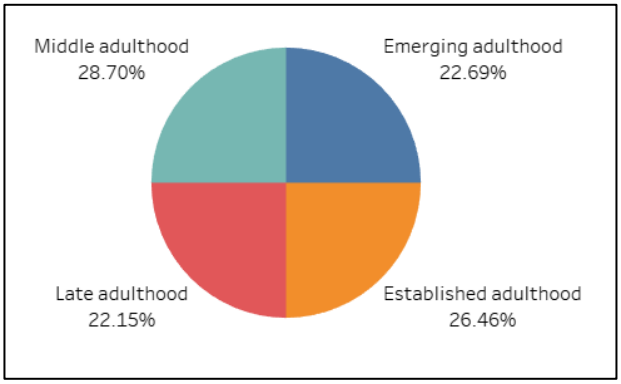


FIGURE 2. Percentage of participants by age group.

The education level of the people who participated in the studies is presented through the following levels: primary education, vocational training, secondary education, postsecondary education, university degree, no education.

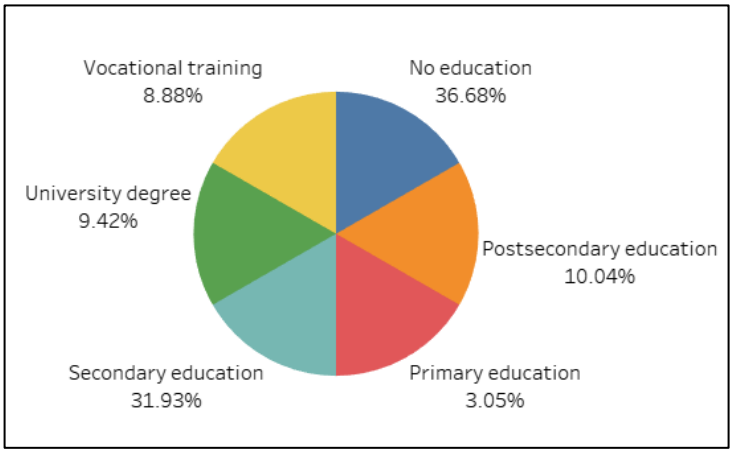


FIGURE 3. Percentage of participants by education level.

Participants were also asked whether, or not, there were any changes regarding the finances and income during the pandemic period. They gave the answers: 1 - has worsened to 7 - has improved. People who participated in the studies had also questions regarding income. They answered as no, yes, or didn't answer at all. Besides financial situation and income continuity, participants were asked whether they suffered from any pre-existing medical conditions that could cause or accelerate COVID-19 infection. They gave answers as no, yes, or don't know. As a part of questionnaires, participants were also asked about their employment status, or whether they are students or not. By checking the Class column, labeled by yes, or no, depending on whether the participant has signs of depression and anxiety, or not, we can see that 50.85% of participants has signs of depression and anxiety, while 49.15% of participants doesn't have any signs of depression and anxiety.

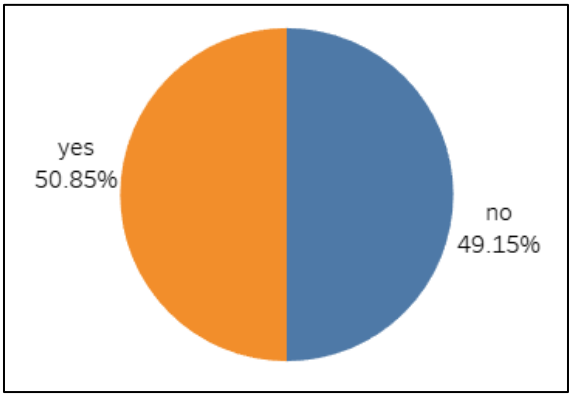


FIGURE 4. Percentage of participants depending on whether a participant has (yes), or has not showed (no) symptoms of depression and anxiety.

C. Preprocessing

During the work on these studies, after aggregation of the data, the next step was to convert them into an acceptable format for any work. Data preprocessing is the method of transforming raw data into a graspable format. Collected data is usually incomplete, noisy, inconsistent, and redundant. Preprocessing data is an important step to reinforce data effectiveness. This method involves numerous steps that facilitate converting raw data into a processed and wise format that is prepared for further work.

The first part of data preprocessing for this study is done in the Jupyter notebook. Data is cleaned by deleting unnecessary columns, and columns with empty fields are filled with a. The second part of the data preprocessing is done in the Excel tool. We were using Excel formulas to calculate values in the target column, which shows whether the particular record has/has not any signs of depression and anxiety, based on the answers to questionnaires. After data is preprocessed, it is prepared for any further training and testing model phases.

D. Machine Learning Methods

After data cleaning, pre-processing, and wrangling, the next step to be done is to feed it to the model and obtain output in probabilities. Since this is a classification problem, classification algorithms are used to predict if participants, aged 18-85, have symptoms of depression and anxiety, or not. Classification rules are determined within the training phase, and also the same doesn't seem to be modified within the validation/testing stage. Classification algorithms used to predict whether or not the participant has symptoms of depression and anxiety are:

- *Decision tree (J48)* is among the most popular decision tree algorithms in machine learning written works. If the decision tree is simply too populated a tree pruning methodology will be accustomed to remove inessential attributes and restart classification operation once more.
- *Logistic Regression* may be a linear algorithm that is straightforward and fast, however can be very effective on some kinds of problems. It works by predicting class probabilities instead of actual classes and uses logic transform to predict probabilities directly. The J48 algorithm for instance uses probabilities internally to assist with pruning. The logistic regression solely supports binary classification issues, although the Weka implementation has been changed to support one-vs-all classification issues.

While working on this research and master thesis, we were using two tools: *Weka*, to test and train the ML models, and *Tableau*, to visualize the data.

4. Results

A. Performance Evaluation

To measure accuracy for machine learning classification, we tended to use a *confusion matrix*. Within the confusion matrix, every row in a table points out the samples of predicted values whereas every column represents the samples of the actual values (or vice versa). Members of the dataset that are classified properly are set on the matrix's diagonal. The accuracy of the algorithms is calculated by the division of the total sum of all elements by a trace of the matrix (sum of the diagonal elements). If we have a case of binary classifier (which is in our case), then we will have only two labels: "Normal" and "Abnormal".

Detailed accuracy by class for all four samples (True Positive, True Negative, False Positive, False Negative) is presented through the following parameters: TPR/Recall, FPR, Precision, F-Score, MCC, ROC Area, PRC Area. The meaning of parameters is presented through the following bullet points.

TABLE 1. Confusion matrix illustrated with two-class system.

	Predicted: Abnormal	**Predicted:** Normal
Actual: Abnormal	True Positive	False Negative
Actual: Normal	False Positive	True Negative

- **TPR/Recall**

True Positive Rate (TPR) is used to measure the percentage of actual positives that are accurately determined as positive ones. TPR is calculated as a division of the total count of accurately determining positive samples (TP) and total count of positive samples (TP + FN):

$$TPR = \frac{TP}{TP + FN}$$

On the other side, *True Negative Rate (TNR)* is an outcome where the model is correctly predicting the negative class (actual negatives which are correctly identified). TNR is calculated

as a division of the total count of accurately determining negative samples (TN) and total count of negative samples (TN + FP):

$$TNR = \frac{TN}{TN + FP}$$

- **FPRconstant value**

False Positive Rate (FPR) is used to measure the percentage of actual negatives which are inaccurately determined as positive samples. FPR is calculated as a division of the total count of negative samples inaccurately determined as positive samples (FP) and total count of negative samples (FP + TN):

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

On the other side, *FNR (False Negative Rate)* is used to measure the percentage of actual positives that are inaccurately determined as negative samples. FNR is calculated as a division of the total count of positive samples inaccurately determined as negative samples (FN) and total count of positive samples (FN + TP):

$$FNR = \frac{FN}{FN + TP}$$

When examining model accuracy, typically two main measures that are considered are TPR and FPR.

- **Precision**

Precision is used to measure how precise, or accurate our model is - out of those predicted positive, how many of them are positive. Precision is calculated as a division of the total count of positive samples accurately determined as positive samples (TP) and total count of predicted positive samples (TP + FP):

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

- **F-Score**

F-Score/ F1-Score is a criterion of a model's accuracy on a dataset. It is used when evaluating binary classification systems, where examples are classified into 'positive', or 'negative'. F-Score is the harmonic mean of the precision and recall:

$$F1\ Score = \frac{Precision * Recall}{Precision + Recall} * 2$$

- **MCC**

Matthews correlation coefficient (MCC) is a coefficient of correlation of the observed about predicted binary classifications. Outcome values are in a range from −1 to +1.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

From this formula for calculating MCC, we can conclude following:

- when FP = FN = 0 (perfect classifier), MCC = 1, indicating perfect positive correlation
- when TP = TN = 0 (classifier is misclassifying), MCC = -1, indicating perfect negative correlation

- **ROC Area**

ROC (Receiver Operator Characteristic) Area, or Area under the ROC (AUC-ROC) serves for measurement of the 2D area under the whole ROC in a range from (0,0) to (1,1). ROC displays the performance of the model at all classification inceptions.

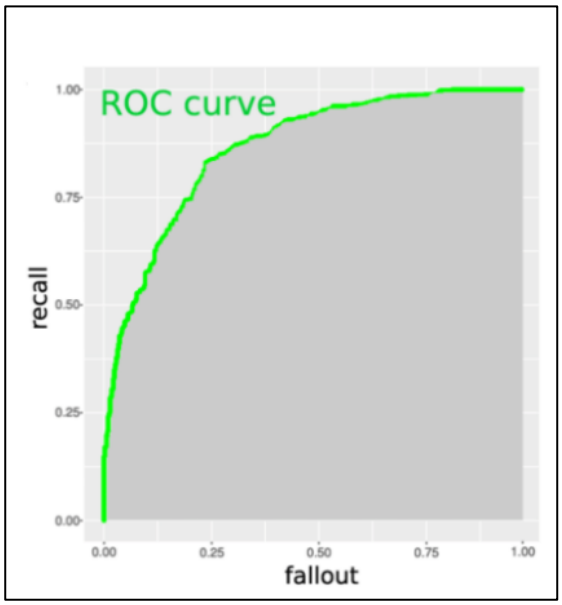


FIGURE 5. The grey area is AUC-ROC (Area under the ROC curve), where the recall (True Positive Rate) score is positioned on the y-axis and the fallout (False Positive Rate) score on the x-axis.

- **PRC Area**

PRC (Precision-Recall Curve) Area, or Area under the PRC (AUC-PRC) is used to measure 2D area under the whole PRC curve in a range from (0,0) to (1,1). The precision-recall curve (PRC) shows precision values for corresponding recall values.

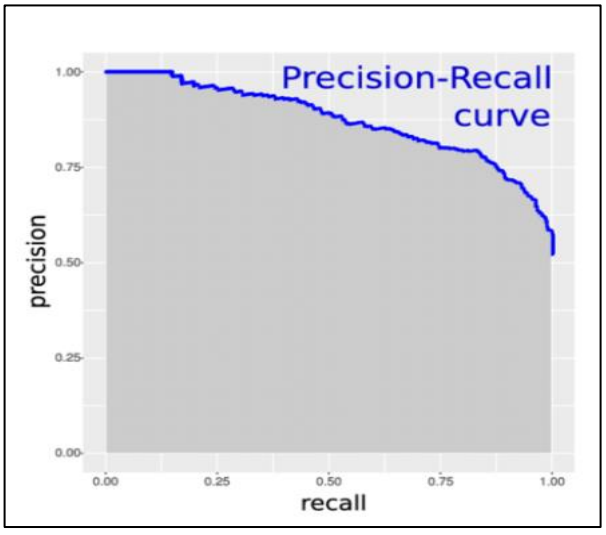


FIGURE 6. The grey area is AUC-PRC (Area under the PRC curve), where the precision score is positioned on the y-axis and the recall score on the x-axis.

As a part of this research, results are tested to determine which method has more acceptable performance and if there is any difference in misclassification rates that is significant for statistics. Algorithms are trained and tested on the common dataset, on the true labels. The method that has better performance is shown as the more acceptable one.

A. Results by Classification algorithms

To obtain classification models results, we tended to use Decision Tree (J48) and Logistic Regression machine learning algorithms, where results were obtained using the Weka tool. Classification models are tested using percentage split choice, where the dataset was split in the ratio of 80:20. *That is 80% of data goes to the training set, 20% to the test set.* It is vital that the data that is being used to train and which is being used to test the model respect as similar statistical distribution as possible, w*hereas calculating performance measures,* precision, recall, and accuracy should be high as possible.
For the decision tree (J48) classification model, the accuracy score results were 86.0987%. Our Decision Tree (J48) model has a Precision of 0.861 - in other words, when our ML model predicts a participant has symptoms of depression and anxiety, it is correct 86.1% of the time. Our Decision Tree (J48) ML model has a Recall of 0.861 - in other words, it correctly identifies 86.1% of all symptoms of depression and anxiety. Additionally, if F1-Score has good value, that would be a sign of a fine Precision and a fine Recall value, too. In the case of our Decision Tree (J48) ML model, F1-Score = 0.861 indicates that both, Precision and Recall have good values. We have the value of 0.824 as the ROC Area (AUC-ROC), which is a good score. In simplest terms, this means that the model will be able to distinguish the participants with the signs of depression and anxiety and those with no symptoms 82% of the time. Just as for AUC-ROC, we got a good PRC Area (AUC-PRC) of around 78%. In Table 3. are interpreted confusion matrix values for the Decision Tree (J48) model, where a=no, b=yes.

TABLE 2. Detailed accuracy by class for Decision Tree (J48).

	TPR	**FPR**	**Precision**	**Recall**	**F-Score**	**MCC**	**ROC Area**	**PRC Area**	**Class**
	0.858	0.137	0.850	0.858	0.854	0.721	0.824	0.762	no
	0.863	0.142	0.871	0.863	0.867	0.721	0.824	0.803	yes
Weighted Avg	0.861	0.139	0.861	0.861	0.861	0.721	0.824	0.783	

TABLE 3. Confusion matrix for Decision Tree (J48) model.

	Predicted (a):	**Predicted (b):**
Actual (a):	91	15
Actual (b):	16	101

For the Logistic Regression classification model, the accuracy score results in 83.4081%, and in Table 4. The accuracy by class for the Logistic Regression classification model is elaborated in detail.

TABLE 4. Detailed accuracy by class for Logistic Regression.

	TPR	FPR	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.877	0.205	0.795	0.877	0.834	0.672	0.903	0.889	no
	0.795	0.123	0.877	0.795	0.834	0.672	0.903	0.924	yes
Weighted Avg	0.834	0.162	0.838	0.834	0.834	0.672	0.903	0.907	

Our Logistic Regression ML model has a Precision of 0.838 - in other words, when our ML model predicts a participant has symptoms of depression and anxiety, it is correct 83.8% of the time. Our Logistic Regression ML model has a Recall of 0.834 - in other words, it correctly identifies 83.4% of all signs of depression and anxiety. In the case of our Logistic Regression ML model, F1-Score = 0.834 indicates that both, Precision and Recall have good values. We get a value of 0.903 as ROC Area (AUC-ROC), which is a fine score. In simplest terms, this means that the model will be able to distinguish the participants with the symptoms of depression and anxiety and those with no symptoms 90% of the time. Just as for AUC-ROC, we got a good PRC Area (AUC-PRC) of around 90%. Table 5. shows confusion matrix values for the Logistic Regression classification model, where a = no, b = yes.

TABLE 5. Confusion matrix for Logistic Regression model.

	Predicted (a):	Predicted (b):
Actual (a):	93	13
Actual (b):	24	93

5. Discussion

This paper presents research based on changes in mental health caused by the irruption of the COVID-19 pandemic. By this, we can investigate more about the changes in a psychological state that people are coming through and find out in which way we can help people with mental disorders. The number of people who suffer from depression and anxiety increased significantly during the pandemic. Analysis of data is done by using ML tools predicting the group of participants with signs of depression and anxiety, based on his/her answers to the questionnaires, and the attributes of the participants. For the measurement of the performance of classification models we were using confusion matrices, and also did the comparison of the accuracy between decision tree and logistic regression model. The analysis reached 86.1 % and 83.4% classification accuracy for Decision Tree and Logistic Regression classification models, respectively.

Based on the results given by the studies, the youngest population (age 18-29), participated in the surveys, experienced more intense depression and anxiety symptoms than participants from other age groups. Results from the other studies showed that people from the oldest age groups more often experience depression and anxiety symptoms, while our studies provided conclusions that the oldest age groups were experiencing the lowest intensity of depression and anxiety signals throughout the COVID-19 period.

Both, our studies and related studies, which were based on the research of the impact of COVID-19 on mental health, found that females are more affected by depression, anxiety, and distress.

Our results show that the population with low income and financial instability were experiencing more severe symptoms of depression and anxiety, as found in previous research for population who experienced job loss and financial struggles. The pandemic period brought people online. This way of communication was more restrictive to the older generations since most of

them are not pretty familiar with the technology. By this, the oldest age group felt abandonment and experienced a higher level of difficulties, related to lack of communication. In the end, lack of live meetings, building relationships, and a sense of abandonment was among the factors for depression and anxiety symptoms appearance between the younger ones (age 18-29).

6. Conclusion

Anxiety and depression symptoms were increasing significantly throughout the lockdown caused by the COVID-19 pandemic. The isolation and insecurity caused by the pandemic have made mental health even more impaired. Various factors that are predicting depression and anxiety symptoms within completely different age teams recommend that support should be distributed based on age since the performance of most tasks that were done daily has changed a lot throughout the COVID-19 period. This means that its performance needs to be adjusted, taking into account people's characteristics. This research process is used to indicate possible measurements, therapies, and solutions that will help those that were exposed to the risk the most. By being vigilant and with the possibility to identify potential threats, we will be prepared to come up with good solutions, to help in avoiding unhealthy outcomes.

7. References

[1] Gambin, M., Sekowski, M., Woźniak-Prus, M., Wnuk, A., Oleksy, T., Cudo, A., ... Maison, D. (2020, June 29). Generalized anxiety and depressive symptoms in various age groups during the COVID-19 lockdown. Specific predictors and differences in symptoms severity. https://doi.org/10.31234/osf.io/42m87

[2] Hafstad, G. S., Sætren, S. S., Wentzel-Larsen, T., & Augusti, E.-M. (2021). Adolescents' symptoms of anxiety and depression before and during the Covid-19 outbreak – A prospective population-based study of teenagers in Norway. The Lancet Regional Health - Europe, 5, 100093. https://doi.org/10.1016/j.lanepe.2021.100093

[3] Bjelland, I., Krokstad, S., Mykletun, A., Dahl, A. A., Tell, G. S., & Tambs, K. (2008). Does a higher educational level protect against anxiety and depression? The HUNT study. *Social science & medicine (1982)*, *66*(6), 1334–1345.

[4] Hammarberg, K., Tran, T., Kirkman, M., & Fisher, J. (2020). Sex and age differences in clinically significant symptoms of depression and anxiety among people in Australia in the first month of COVID-19 restrictions: A national survey. BMJ Open, 10(11), e042696. https://doi.org/10.1136/bmjopen-2020-042696

[5] Single Care Team, Anxiety statistics, https://www.singlecare.com/blog/news/anxiety-statistics/, 2021.

[6] Nochaiwong, S., Ruengorn, C., Thavorn, K., Hutton, B., Awiphan, R., Phosuya, C., Ruanta, Y., Wongpakaran, N., & Wongpakaran, T. (2021). Global prevalence of mental health issues among the general population during the coronavirus disease-2019 pandemic: A systematic review and meta-analysis. Scientific Reports, 11(1), 10173. https://doi.org/10.1038/s41598-021-89700-8

[7] Delmastro, M., & Zamariola, G. (2020). Depressive symptoms in response to COVID-19 and lockdown: A cross-sectional study on the Italian population. Scientific Reports, 10(1), 22457. https://doi.org/10.1038/s41598-020-79850-6

[8] Kamran Ul haq, A., Khattak, A., Jamil, N., Naeem, M. A., & Mirza, F. (2020). Data Analytics in Mental Healthcare. Scientific Programming, 2020, 1–9. https://doi.org/10.1155/2020/2024160

[9] Stewart, R., & Davis, K. (2016). 'Big data' in mental health research: Current status and emerging possibilities. Social Psychiatry and Psychiatric Epidemiology, 51(8), 1055–1072. https://doi.org/10.1007/s00127-016-1266-8

Data Sharing

The dataset is administered by the *University of Warsaw, Social Sciences*. Dataset has been downloaded from the *Harvard Dataverse* site (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0NP102).

Prevalence of es2108622 (*CYP4F2*3*) Single Nucleotide Polymorphism – A Review

Selma Žiga*, Hana Efendić*, Larisa Bešić*
[1]International Burch University,
selma.ziga@stu.ibu.edu.ba
hana.efendic@stu.ibu.edu.ba
larisa.besic@ibu.edu.ba

Literature review

Abstract: *Cardiovascular diseases are known to be treated with anticoagulants lifelong. Warfarin is one of the most commonly used medications for anticoagulation despite causing serious side effects in some patients. Different single nucleotide polymorphisms (SNPs) that have a role in the cytochrome P450 system can also affect the metabolism, as well as dosing, of warfarin. The purpose of this review is to look into the prevalence of this SNP in the past research and screen for possible correlations with age, place of origin, family history of cardiovascular and cerebrovascular diseases, or other medical conditions possibly present in various populations. In total, 20 scientific articles falling under the inclusion criteria were reviewed and found usable, and the rest of the cases will be highly beneficial in the upcoming years to determine the role of the recently discovered CYP4F2 rs2108622 variant, as well as the previously known CYP2C9 and VKORC1 SNPs, in the variance of warfarin dose requirement. These findings may also point researchers in the right direction for qualifying and validating these genetic variants for use as GBs (genomic biomarkers) in the clinical and medical practice of treatment with warfarin.*

Keywords: SNP, rs210862 (*CYP4F2*3*) SNP, warfarin, vitamin K, cardiovascular disease, anticoagulation, INR.

1. Introduction

Throughout the patient's life, cardiovascular diseases are treated with anticoagulants. These diseases include transient ischemic attack (TIA), coronary artery bypass grafting (CABG), pulmonary thromboendarterectomy (PTE), rheumatic heart disease (RHD), atrial/aortic valve replacement (AVR), permanent pacemaker (PPM), percutaneous coronary intervention (PCI), cerebrovascular accident (CVA), left atrial (LA) clot, mitral valve replacement (MVR), atrial fibrillation (AF), deep vein thrombosis (DVT), percutaneous transvenous mitral commissurotomy (PTMC) as well as combined conditions. Anticoagulants interact at various stages of the coagulation cascade and are classified into two types: those that act directly by inhibiting enzymes and those that act indirectly by binding to antithrombin or preventing its synthesis in the liver. [1]. According to the types of anticoagulants, there are several available: low molecular weight heparin (LMWH), unfractionated heparin (UFH), low molecular weight heparin (LMWH), direct factor 10a Inhibitors and vitamin K-dependent antagonists, and direct thrombin inhibitors. Medical conditions, pat preferences, and risk stratification should be taken into consideration when choosing the appropriate anticoagulant. Atrial fibrillation, venous thromboembolism, and post-heart valve replacement are the most common medical conditions requiring anticoagulation treatment. Venous thromboembolism is important because it is often one of the first symptoms of several other medical conditions [2]. Warfarin, a vitamin K-dependent antagonist, is one of the most commonly used anticoagulants. Warfarin works by inhibiting the enzyme vitamin K epoxide reductase which is required for the gamma-carboxylation of vitamin K-dependent factors. The dosing limit is low, and the effect is highly influenced by various factors (including diet), which may lead to resistance to the treatment. Treatment with these anticoagulants requires regular monitoring with an International Normalized Ratio. This enables the usage of a standardized method of analysis and reporting the consequences of an oral anticoagulant (like warfarin) intake, specifically related to blood clotting. Results will vary according to medication the patient takes, age, and any additional health issues. The INR number should be between 2 and 3 if a patient takes an anticoagulant, but it could be different, depending on the patient's condition. Single nucleotide polymorphisms are recognized as the most popular molecular markers for genetic studies as they are a type of variation of a single base pair of polymorphism. SNPs have been found to associate with drug response, diseases, and other phenotypes [3].

*A. Variant rs2108622 of CYP4F2*3*

CYP4F2 is a cytochrome P450 enzyme. It participates in the -hydroxylation of arachidonic acid and vitamin E. Researchers discovered that CYP4F2 participates in VK1 metabolism, and that the rs2108622 polymorphism may affect VK1 oxidase activity. When investigating VK1 oxidase activity in human liver microsomes, the *CYP4F2* CC pool showed to have the highest activity. On the other hand, the *CYP4F2* TT pool demonstrated a decrease of 75%, and the CT pool had its activity on the intermediate level. The reason behind this is that the carriers of rs2108622 might have higher levels of VK1 oxidase, thus requiring a higher dose of warfarin. The rs2108622 polymorphism has been found to impact warfarin dose necessity and clarify roughly 2% - 7% of the variance. *CYP4F2* was also found to be a minor predictive of medication dose in a genome-wide association study (GWAS) involving 1,053 Swedish subjects. Those certain research results revealed that *CYP4F2* rs2108622 T carriers need to have a higher warfarin dose and that *CYP4F2* could be the 3rd hereditary predictive of warfarin daily dosage [4].

Studies chosen for this review are based on the connection of the rs2108622 single nucleotide polymorphism with warfarin dose, the reactions performed within the patient, and INR in populations. The goal is to investigate the prevalence of this single nucleotide polymorphism in previously published studies and detect possible correlations with age, place of origin, family history of cardiovascular and cerebrovascular diseases, or other medical conditions. It is very important to take all of these into consideration during diagnosis, as well as treatment.

2. Materials and methods

A. *Search Strategy*

Databases used in this review were BioMed Central (BMC), National Center for Biotechnology Information (NCBI), Journal of Human Genetics, PLoS Genetics, Future Medicine, and Science Direct. These databases were used due to a large number of free articles, and also a higher number of articles related to the topic. The search was firstly based on keywords related to the topic, namely rs2108622 SNP, *CYP4F2*, PCR, Warfarin, Vitamin K, Anticoagulation, INR, Cardiovascular diseases, and any combination of these.

B. *Inclusion and Exclusion Criteria*

Table 1 shows the inclusion and exclusion criteria used in this paper. Only studies written in the English language were included to avoid translational mistakes. Articles that were not dealing with previously mentioned keywords were excluded. Sources were manually and thoroughly analyzed to exclude duplicates and potentially biased articles.

TABLE 1. Inclusion and exclusion criteria.

Inclusion criteria	**Exclusion criteria**
Language: English	Any other language
Focusing on *CYP4F2* rs2108622, warfarin treatment and cardiovascular diseases, and connection between these (and other keywords listed above)	Papers incompletely or not include any of the selected keywords, especially *CYP4F2* rs2108622, warfarin treatment, and cardiovascular diseases
Peer-reviewed papers	Unreviewed articles, textbooks

3. Results and Discussion

Before implementing the inclusion and exclusion criteria, 34 studies were found to be in line with the topic according to keywords, but four were immediately excluded because they were inaccessible.

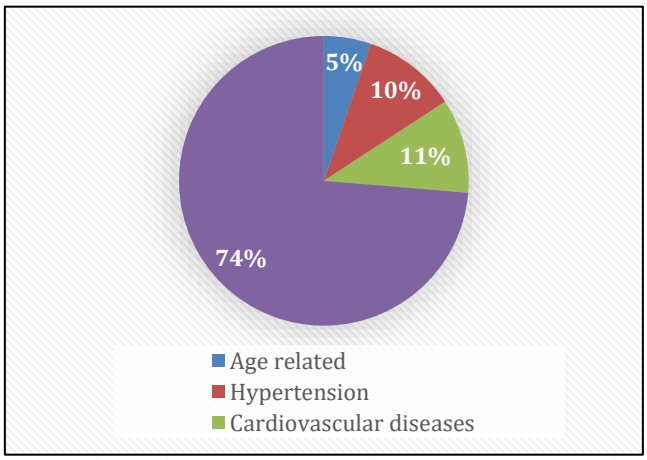


FIGURE 1. Summary of the number of papers chosen for the review based on the topic the studies were focusing on. It outlines the general organization of the review and findings.

After exclusion criteria were applied, five articles were excluded because they were published by unapproved sites (not scientific journals), and five were excluded because they were not based on warfarin maintenance or rs2108622 SNP. In total, 20 scientific articles falling under inclusion criteria were reviewed, meaning those published in the English language, and focusing on warfarin dosing in correlation with age, cardiovascular diseases, hypertension, and INR (Sakiene *et al*, 2016; Pawlowska *et al*, 2019; Karpinos *et al*, 2013; Bener *et al*, 2013; Zhang *et al*, 2017; Geng *et al*, 2019; Luo *et al*, 2015; Khosropanah *et al*, 2017; Banecka-Majkutewicz *et al*, 2012; Meng *et al*, 2015; Liao *et al*, 2016; Li *et al*, 2018; Li *et al*, 2012; Ross *et al*, 2010; Kumar *et al*, 2014; Sipeky *et al*, 2015; Krajčíová *et al*, 2014; Borgiani *et al*, 2009; Caldwell *et al*, 2008; and Takeuchi *et al*, 2009). Out of these, fourteen were related to warfarin dosing in populations, three of which included data on patients with cardiovascular diseases. Two papers were related to cardiovascular diseases, two with hypertension and one with age (Figure 1).

A. rs2108622 Single-Gene Polymorphism (CYP4F2) – Age-Related

Neurodegenerative disorder named age-related macular degeneration (AMD) seems to be a disorder that is the main cause of permanent blindness in people over the age of 65, especially in Western countries [5]. The amount of people with AMD is estimated to rise by roughly 50% by 2020, and the disease's responsibility is expected to increase with age. According to Pawlowska et al. (2019), the accumulation of oxidized lipids appears to play a central role in the growth of AMD [6].

Sakiene *et al* (2016) analyzed patients with exudative age-related molecular degradation and patients with early age-related macular degeneration. The experiment was done by DNA extraction (blood) and PCR reaction. The comparison of the rs2108622 genotype frequency by age groups did not reveal significant differences. The comparison between male and female carriers of the rs2108622 (people who have to die to diagnose AMD and the control group) did not reveal any significant differences, but males and females with AMD did. The analysis showed that the codominant variables inside the group of persons under the age of 65 were significant statistically. It is conceivable that gene-environment interactions influenced the genotype distribution of rs2108622 in patients with exudative AMD and control subjects [5].

B. Correlation CYP4F2 (rs2108622) Gene Polymorphism and Hypertension

The most common cardiovascular disease is hypertension [7]. Hypertension has recently been identified as a complex multifactorial illness caused by interactions between countless genetics and the environment [8], [9]. A study found that rs2108622, rs1558139 and rs2108622 single nucleotide polymorphisms on the gene *CYP4F2* are linked to hypertension, with the rs1558139 polymorphism being extremely powerful in males. These findings are also supported by six studies, three of which looked into the rs1558139 polymorphism and six of which looked into the rs2108622 polymorphism [10].

Luo *et al* (2015) included 4 studies that contain a total of 1878 patients with hypertension and 1512 healthy control subjects. The study, which contained 4 independent case-control studies, has shown that the *CYP4F2* gene rs2108622 polymorphism was not linked to an increased risk of high blood pressure [11].

C. Correlation Between Cardiovascular Diseases and CYP4F2 Gene Rs2108622 Polymorphism

Warfarin has a wide range of applications, including pulmonary embolism, stroke, and preventative measures of thromboembolic events, as well as atrial fibrillation, coronary dysfunction, and prosthetic valve placement. Due to the FDA's Adverse Event Reporting System, warfarin has been one of the 10 leading drugs with the most serious side effects, particularly

during the initial phase of treatment [12]. Figure 2 depicts the cardiovascular diseases included in this study: ischemic stroke, cardiac ablation, valve replacement, and cardiovascular patients in general.

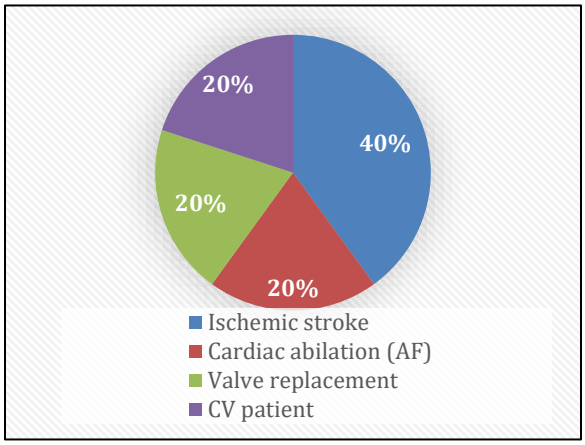


FIGURE 2. Summary of the number of CV disease-related papers chosen for the review based on the type of CV disease the studies were dealing with.

The figure outlines the general organization of one subtopic discussed in the review. Ischemic stroke is caused by the interaction of genetics and the environment, and many genetic factors can affect its pathogenesis. Of 100% possibility to have any of strokes it is 80% to have Ischemic stroke (IS) [13].

Six studies were performed and included a total of 2,187 cases and 7,556 controls, and the correlation between *CYP4F2* V433M polymorphisms and susceptibility to ischemic stroke was evaluated truly by the Recessive model. The methodology they used is Q statistic test methods and I^2 quantitative assessment for evaluating the heterogeneity size between the studies. On the one hand, the study can only combine the crude OR values to examine the association because it is impossible to collect the OR (odds ratio) value of each study after adjusting for age and sex. As a result, there may be a discrepancy between the test and real-life association intensity. However, due to the small number of papers included, it is not possible to analyze the connection in the subgroup by race [14].

Another study took 396 patients with ischemic stroke and 378 controls were genotyped for rs9333025, rs3093135, rs2269231, and rs2108622. Generalized multifactor dimensionality reduction (GMDR) methodologies were used to investigate gene-gen interactions. The GMDR analysis revealed that rs2108622 and rs9333025 had a strong gene–gene interaction. Polymorphisms: rs9333025 GG and rs2108622 GG genotypes, this gene–gene interaction predicted a significantly higher risk of ischemic stroke [15].

One of the most represented heard conditions is atrial fibrillation (AF). Research has been conducted within the Chinese population in comparison with African and Caucasian populations. According to the 2010 Chinese Census, China's population of AF cases aged 35 years is 5.26 million, and the number of AF ablation procedures is quickly increasing. AF catheter ablation is a surgery that necessarily requires rest for three weeks, and warfarin therapy for three months. Warfarin dosage is presented with the INR and carefully regulated due to ethnic and individual dosing variances.

Jiao Li *et al* used a total of 222 patients from West China Hospital in their study (82 males and 140 females). Their demographic data, such as body weight, constituent warfarin dosage, age, height, was meticulously documented, and blood samples were taken for DNA isolation. Methodology they used was DNA extraction and PCR reaction. The genotyping results performed

low frequency of variant rs2108622 *CYP4F2* gene of all T allele carriers. The TT and CT carriers required a significantly higher dose of warfarin in *CYP4F2* rs2108622 genotype patients. Sichuan Chinese patients were present a low frequency of warfarin dose which can be the main cause of sensitivity to warfarin and their requirements in warfarin dosage in compared with Caucasians [16].

Li *et al* (2012) studied 352 patients after heart valve replacement surgery Patients' warfarin dosing was considered to obtain an INR of 1.8 to 2.5. They investigated for SNPs in *CYP4F2* in these patients and looked into their relationship with warfarin dosing. The following information was collected for each patient: (1) general information, (2) drug, (3) surgical history, (4) medical history. DNA samples deployed were isolated from blood. Methodology conducted involved DNA extraction and PCR. The study also includes 352 patients, 228 of these have a CYP4F2 wild-type homozygous CC genotype, 104 had a heterozygous CT genotype, and 20 had a mutant TT genotype. The results show that the warfarin dose necessity increased substantially in Chinese people who have at least one T allele versus those who are homozygous for the C allele [17].

In the study performed by Khosropanah et al. (2017), 226 cases were selected from the 230 participants, with 152 subjects classified into group A (dosage of warfarin is 5 mg/day) and 74 cases classified into group B (dosage of warfarin is above 5 mg/day). The study's findings show that when Iranians have at least one T allele, their warfarin dose requirement increases significantly when compared to those who are homozygous for the C allele [12]. One of the limitations of such studies involving a wide range of diseases, duration of warfarin therapy, and age of patients participating in the study is that it may have to affect statistical analysis methods.

D. Warfarin Dosage Response Related Pharmacogenetics in populations

Because there is little possibility of over-or under-anticoagulation warfarin dosing is to hold prothrombin time 2 – 3 of the international normalized ratio (INR). Earlier studies have shown that two genes have interacted with the vitamin K-dependent clotting pathway, *VKORC1* and *CYP2C9,* account for an additional 30–54 percent of the variant in dosing warfarin [18]. South Indians, Roma and Hungarians, Slovaks, Chinese, Italians, Caucasians, Asians, Africa, Swedish, and Koreans were all studied.

Kumar *et al* (2014) reported the mean daily requirement dose of warfarin to be 4.7 ± 2.1 mg/day in the South Indian population with increased warfarin therapy in patients with rs2108622 SNP [19].

Sipeky *et al* (2015) found that members of the Roma population have an elevated chance for higher mean warfarin dose requirement in comparison with the Hungarian population, besides a decreased risk of major bleeding events in long-term warfarin use [20]. Another study showed that polymorphisms in the genes CYP4F2 have a smaller effect on warfarin pharmacogenetics than in VKORC1 and *CYP2C9* and which have a meaningful impact on personal reaction to warfarin dosages in the Slovak population [21]. Lastly, scientists reported warfarin maintenance in the Italian population. Consumers included just patients with CV disorders, and it is clear from these observational studies that the quantity of required warfarin dose modifications was reduced, and patients were outside of the target INR range [22]. The possible medical advantage of CYP4F2 genotyping differs by the racial group due to the difference in the frequency of the foundational gene variants between many substantial racial groups. The minimal allele frequency for CYP4F2 is roughly 30% in Asians and Caucasians while being only 7% in African populations. As a result, the African population's tends stable dose of warfarin is expected to be lower than that of Asians and Caucasians [23].

Takeuchi *et al* (2009) strongly indicate that the 3rd gene, *CYP4F2* (rs2108622) affects warfarin dose in the Swedish population, with an INR of 3.0-4.0 [24]. According to Ross *et al* (2010), gender, age, INR, BSA, and *CYP2C9, VKORC1,* and *CYP4F2* polymorphisms all influence warfarin dosage requirements in the Korean population. The most recent variation suggested affecting warfarin dosing was also knowns as a non-synonymous - rs2108622 SNP (C/T polymorphism) located within the *CYP4F2* gene. A recent genome-wide association (GWA) study

on patients in the Swedish population examined the relationship of rs2108622 with a dosage of warfarin after controlling for the effects of *CYP2C9* and *VKORC1* [18].

After analyzing the studies, it can be concluded that the rs2108622 SNP is one of the three main factors for determining the dosage of warfarin, which may be detected with enzyme restriction. DNA source was blood for all the studies and considering the patient's conditions, just one group of investigators have included a healthy positive control group, while others included patients with cardiovascular disorders. It is of utmost importance to refer to a control group in order to "calibrate" the results, and probably lead to more comprehensive conclusions and possible comparisons. On the other hand, the information obtained and organized in this review increases comprehensiveness.

4. Conclusion

Usable cases will be incredibly beneficial in the coming few years to determine the role of the recently discovered *CYP4F2* rs2108622 single nucleotide polymorphism, as well as the previously known *CYP2C9* and *VKORC1* SNPs, in the variance of requirement dose of warfarin. The findings of these studies may also point researchers in the right direction for qualifying and validating these genetic variants for use as GBs (genomic biomarkers) in the clinical and medical practice of treatment with warfarin.

5. References

[1] Momodu, I. I. (2020). Anticoagulation. *StatPearls [Internet]*.

[2] Schein, J. R., White, C. M., Nelson, W. W., Kluger, J., Mearns, E. S., & Coleman, C. I. (2016). Vitamin K antagonist use: evidence of the difficulty of achieving and maintaining target INR range and subsequent consequences. *Thrombosis Journal*, *14*(1), 1-10.

[3] Dolan, G., Smith, L. A., Collins, S., & Plumb, J. M. (2008). Effect of setting, monitoring intensity and patient experience on anticoagulation control: a systematic review and meta-analysis of the literature. *Current medical research and opinion*, *24*(5), 1459-1472.

[4] Cen, H. J., Zeng, W. T., Leng, X. Y., Huang, M., Chen, X., Li, J. L., ... & Zhao, L. Z. (2010). CYP4F2 rs2108622: a minor significant genetic factor of warfarin dose in Han Chinese patients with mechanical heart valve replacement. *British journal of clinical pharmacology*, *70*(2), 234-240.

[5] Sakiene, R., Vilkeviciute, A., Kriauciuniene, L., Balciuniene, V. J., Buteikiene, D., Miniauskiene, G., & Liutkeviciene, R. (2016). CYP4F2 (rs2108622) gene polymorphism association with age-related macular degeneration. *Advances in medicine*, *2016*.

[6] [6] Pawlowska, E., Szczepanska, J., Koskela, A., Kaarniranta, K., & Blasiak, J. (2019). Dietary polyphenols in age-related macular degeneration: protection against oxidative stress and beyond. *Oxidative medicine and cellular longevity*, *2019*.

[7] Karpinos, A. R., Roumie, C. L., Nian, H., Diamond, A. B., & Rothman, R. L. (2013). High prevalence of hypertension among collegiate football athletes. *Circulation: Cardiovascular Quality and Outcomes*, *6*(6), 716-723.

[8] Bener, A., Darwish, S., Al-Hamaq, A. O., Mohammad, R. M., & Yousafzai, M. T. (2013). Association of PPARγ2 gene variant Pro12Ala polymorphism with hypertension and obesity in the aboriginal Qatari population known for being consanguineous. *The application of clinical genetics*, *6*, 103.

[9] Zhang, J. E., Klein, K., Jorgensen, A. L., Francis, B., Alfirevic, A., Bourgeois, S., ... & Pirmohamed, M. (2017). Effect of genetic variability in the CYP4F2, CYP4F11, and CYP4F12 genes on liver mRNA levels and warfarin response. *Frontiers in pharmacology*, *8*, 323.

[10] Geng, H., Li, B., Wang, Y., & Wang, L. (2019). Association between the CYP4F2 gene rs1558139 and rs2108622 polymorphisms and hypertension: A meta-analysis. *Genetic testing and molecular biomarkers*, *23*(5), 342-347.

[11] Luo, X. H., Li, G. R., & Li, H. Y. (2015). Association of the CYP4F2 rs2108622 genetic polymorphism with hypertension: a meta-analysis. *Genet Mol Res*, *14*(4), 15133-15139.

[12] Khosropanah, S., Faraji, S. N., Habibi, H., Yavarian, M., Mansoori, R., & Haghpanah, S. (2017). Correlation between rs2108622 locus of CYP4F2 gene single nucleotide polymorphism and warfarin dosage in Iranian cardiovascular patients. *Iranian journal of pharmaceutical research: IJPR*, *16*(3), 1238.

[13] Banecka-Majkutewicz, Z., Sawuła, W., Kadziński, L., Węgrzyn, A., & Banecki, B. (2012). Homocysteine, heat shock proteins, genistein and vitamins in ischemic stroke--pathogenic and therapeutic implications. *Acta Biochimica Polonica*, *59*(4).

[14] Meng, C., Wang, J., Ge, W. N., Tang, S. C., & Xu, G. M. (2015). Correlation between CYP4F2 gene rs2108622 polymorphism and susceptibility to ischemic stroke. *International journal of clinical and experimental medicine*, *8*(9), 16122.

[15] Liao, D., Yi, X., Zhang, B., Zhou, Q., & Lin, J. (2016). Interaction between CYP4F2 rs2108622 and CPY4A11 rs9333025 variants is significantly correlated with susceptibility to ischemic stroke and 20-Hydroxyeicosatetraenoic acid level. *Genetic testing and molecular biomarkers*, *20*(5), 223-228.

[16] Li, J., Yang, W., Xie, Z., Yu, K., Chen, Y., & Cui, K. (2018). Impact of VKORC1, CYP4F2 and NQO1 gene variants on warfarin dose requirement in Han Chinese patients with catheter ablation for atrial fibrillation. *BMC cardiovascular disorders*, *18*(1), 1-6.

[17] Li, J. H., Ma, G. G., Zhu, S. Q., Yan, H., Wu, Y. B., & Xu, J. J. (2012). Correlation between single nucleotide polymorphisms in CYP4F2 and warfarin dosing in Chinese valve replacement patients. *Journal of cardiothoracic surgery*, *7*(1), 1-7.

[18] Ross, K. A., Bigham, A. W., Edwards, M., Gozdzik, A., Suarez-Kurtz, G., & Parra, E. J. (2010). Worldwide allele frequency distribution of four polymorphisms associated with warfarin dose requirements. *Journal of human genetics*, *55*(9), 582-589.

[19] Kumar, D. K., Shewade, D. G., Loriot, M. A., Beaune, P., Balachander, J., Chandran, B. S., & Adithan, C. (2014). Effect of CYP2C9, VKORC1, CYP4F2 and GGCX genetic variants on warfarin maintenance dose and explicating a new pharmacogenetic algorithm in South Indian population. *European journal of clinical pharmacology*, *70*(1), 47-56.

[20] Sipeky, C., Weber, A., Melegh, B. I., Matyas, P., Janicsek, I., Szalai, R., ... & Melegh, B. (2015). Interethnic variability of CYP4F2 (V433M) in admixed population of Roma and Hungarians. *Environmental toxicology and pharmacology*, *40*(1), 280-283.

[21] Krajčíová, Ľ., Petrovič, R., Déžiová, Ľ., Chandoga, J., & Turčáni, P. (2014). Frequency of selected single nucleotide polymorphisms influencing the warfarin pharmacogenetics in the S lovak population. *European journal of hematology*, *93*(4), 320-328.

[22] Borgiani, P., Ciccacci, C., Forte, V., Sirianni, E., Novelli, L., Bramanti, P., & Novelli, G. (2009). CYP4F2 genetic variant (rs2108622) significantly contributes to warfarin dosing variability in the Italian population.

[23]Caldwell, M. D., Awad, T., Johnson, J. A., Gage, B. F., Falkowski, M., Gardina, P., ... & Burmester, J. K. (2008). CYP4F2 genetic variant alters required warfarin dose. *Blood, The Journal of the American Society of Hematology*, *111*(8), 4106-4112.

Peer Reviewers

Assist. Prof. Dr. Elnur Tahirović
Assoc. Prof. Dr. Mirza Šarić
Assist. Prof. Dr. Zerina Mašetić
Bećir Isaković, MA
Assoc. Prof. Dr. Monia Avdić,
Assist. Prof. Dr. Saida Sulatnić
Sabina Halilović, MA
Assoc. Prof. Dr. Jasna Hifziefendić
Assist. Prof. Dr. Larisa Bešić
Dželila Mehanović, MA
Assoc. Prof. Dr. Dejan Jokić
Mehrija Hasičić, MA
Assist. Prof. Dr. Dino Kečo

Call for new submission

Background

JONSAE provides a platform for the researchers, academicians, professionals, practitioners and students to impart and share knowledge in the form of high quality empirical and theoretical research papers. The journal covers all areas of Genetics and Bioengineering, Electrical and Electronics Engineering, Information Technology, Architecture, Applied Mathematics, Computer Sciences and Civil Engineering.

Submission and review process

All submissions should be made to email jonsae@ibu.edu.ba.

Submissions must adhere to the format and style Guidelines for JONSAE articles available on the journal's web page. The official referencing style is APA.

Submissions will be subject to an initial screening by our editors and papers that fall outside the scope or which are considered unlikely to be suitable for the JONSAE issue will be desk rejected.

Accepted papers will undergo a typical double-blind review process.
Deadline for submission is 15th of July 2022.

Types of Submission

We welcome high-quality submissions which advance our knowledge on the abovementioned topics. We do not favors any special theoretical perspectives or methodological approaches. The types of acceptable submissions include, but are not limited to:

Theoretical and empirical papers
Literature reviews
Practice reviews
Qualitative, quantitative, mixed-methods research
Experimental research
Single, multiple, large-sample case studies

For any questions, please contact us at jonsae@ibu.edu.ba or publication.office@ibu.edu.ba.