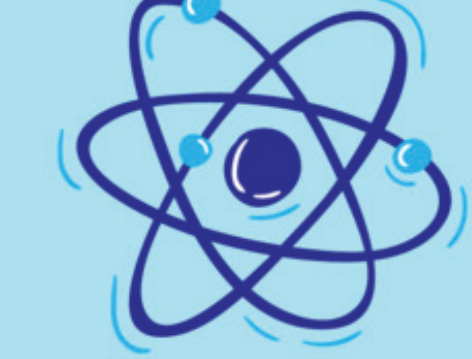
PUBLISHED BY STEMY









A STUDENT RESEARCH JOURNAL

INNOVATION

VOL. 4

Contents

Credits & Acknowledgements

Editor-in-Chief

Gokul Achaththekoot

Graphic Design

Marjorie Mays

Editors

Yaara Aleissa
Mayukha Bhamidipati
Shubh Gupta
Aditi Kona
Sophie Ma

Authors

Gia Mendiratta
Taylor Zekoski
Gokul Achaththekoot
Jonah Xu
Khushi Pola
Sai Jeyaprakash Reddy
Saif Gopang
Salman Shahzad
Shridha Rajeswar
Shubh Gupta
Sonia Nagpal
Spandana Pavuluri
Varun Chandrashekhar
Vedika Venkataramanan

Letter from the Editors

We are STEMY (STEM + Youth): a student-run non-profit organization dedicated to breaking gender, socioeconomic, & racial barriers in STEM education across Louisville. Through our multiple workshop programs & initiatives, we've directly impacted the lives of over 1500 students & will continue to spark a passion for STEM among disadvantaged and underrepresented students.

We began as an after-school club at duPont Manual High School that aimed to help fellow peers with the science fair. However, after about a year, we saw that outside of our own high school, many other schools & programs within Louisville did not receive as many resources & access to STEM education. So, in 2017, we expanded into a 501(c)(3) non-profit working to overcome these barriers.

Effective STEM education is not just valuable to improve academic performance—the right kind of learning influences students' career options, worldview, and mindset. High-quality STEM experiences have been shown to improve problem-solving skills, build confidence, and motivate students to enter high-paying jobs in the future. STEM programs like ours lead participants to view the world through an observation-based, curiosity-driven lens and have the power to open minds that have been closed by years of learning through rote memorization.

We do this through multiple programs, initiatives, & workshops to reach as wide of an audience as we can. Innovation is one of our programs that aims to uplift high school student research. Our two prongs of Innovation are the Peer Mentorship Program & the Innovation Journal. The Peer Mentorship Program, renewed just this year, aims at connecting distinguished upperclassmen highly experienced with student research (such as placing or advancing to the state, national, or international level) to inexperienced student researchers. Mentors help mentees get exposure to the process by reading their papers, contacting labs/professors, and giving quick tips and advice before Science Fair Day. This program typically runs from September to March and from there, we begin the production of our annual, peer-reviewed Innovation Journal. In fact, some of the papers featured in this journal are products of the Peer Mentorship Program. This journal's purpose is to showcase exceptional student research & provide inspiration for others to pursue their own investigation.

Louisville itself provides multiple opportunities for the world of student research. Kentucky itself contests a strong culture of research, sending more representatives in ISEF than most other states. LRSEF, the regional science fair, and DMHSRF, duPont Manual's, work to show how student research is an important aspect in today's STEM-oriented world. We highly encourage you to try it yourself and to reach for the stars!

Innovation, just like our organization, is created and produced entirely by students. We cultivate strong partnerships among editors and authors and give writers the unique opportunity to publish high-quality, peer-reviewed work as a high schooler. Through this journal, we hope to spread our author's innovative spirits and phenomenal research throughout the community. We believe that Innovation is a unique way to expose Louisville to STEM, and we will continue to produce and publish this journal for years to come. Enjoy reading!

Sincerely,
The STEMY Innovation Team

Paving the Path: Why Student Research

The opportunity to conduct research is one of the most valuable experiences for a high school student. It serves as a chance to work with research mentors, shadow experienced professionals in a field of interest, and learn innovative techniques. All it takes is asking a simple question or identifying a real-world problem, which can then culminate into an extensive project that often ends in a proposed solution. In fact, some of the papers published in this issue feature a lab setting; however, due to the nature of the pandemic, some of the most novel projects were done right at home. These home setting projects varied from culturing bacteria on the kitchen counter, compiling EMG signals on the computer, or physically building a listening device. Spandana Pavuluri, a featured author, describes how she adapted to the change. "In the midst of the pandemic, a lot of my original ideas for my project were pushed aside and I had to find something safe that I could do at home." Regardless of the environment, these students have brainstormed innovative ideas and transformed them into final products that progress their field of study.

The process of research is a recursive one, that continues to pose new gaps for student researchers to tackle. One of the most influential aspects of student research is the opportunity to investigate an issue that impacts the local community. This issue features a variety of student research reaching large communities from self machine learning in prosthetics to creating an algorithm for paralysis with applications in conditions such as stroke, ALS, and cerebral palsy. For instance, Shubh Gupta, a featured author, has been investigating the Traveling Salesman Problem, which aims to find the shortest path through a particular set of points — a problem with no general solution. However, after years of research, Shubh developed an estimation algorithm which is beneficial to "companies that deliver goods and are projecting the future, since they can get an estimation for the total distance without knowing the exact locations of the drop off locations." Shubh describes his ultimate goal: "After reading my paper, I hope that people can recognize how seemingly abstract mathematical problems can have direct real world applications." By further exploring them, in areas unknown to many, we can make our world more productive, efficient, and leave it a better place for all.

Moreover, research provides an opportunity for students to pivot towards issues impacting underrepresented communities. They are able to highlight issues that have impacted their own family or friends. This year, STEMY Innovation brought back their Peer Mentorship Program that provided inexperienced high school research students with award-winning student mentors to assist in their research inquiries. The Innovation Journal tackles the same initiatives as the Peer Mentorship Program with the purpose to showcase the outstanding work coming from underrepresented students across Louisville. Amidst the pandemic, Gokul Achaththekoot describes how he came up with an idea by simply noticing an everyday occurrence. "I came up with my project after constantly using disinfecting wipes to clean when guests came over, and I remember that it always left a super strong smell. I wanted to see if there were any health effects or how to see if it's dangerous using these products at home." Through independent research, students take a step towards implementing change in an area that has been inequitable towards serving their best needs. In higher level education, research is especially important for students pursuing STEM careers and yet there is a large disproportion of representation in research for many students coming from different backgrounds. Furthermore, student research is incredibly hard to publish without access to large laboratories and connections in academia. This makes the benefits of early research experience all the more important.

Undoubtedly, student research is a step towards a STEM-oriented career. It not only

Paving the Path: Why Student Research, Cont.

influences a student's future direction, but also helps one achieve goals through practicing hard and soft skills, such as academic writing and time management. Students are able to make useful connections with mentors and present their findings to experts in their field of study. Future employers often seek out prepared and resilient learners with previous experience in intensive problem-solving. As expressed by STEMY alumnus Eddy Zhong, a current sophomore at Yale, "If you do research in high school, that research mindset will already be second nature." The impact of research is seen not only in its ability to strengthen the techniques of students, but also in allowing them to independently lead and realize the impact their work has beyond themselves. Recalling his own research experience, Zhong explains that "doing research helped me develop maturity as a person: any good research scientist knows that you will constantly face setbacks in your experiments, even if you've put in a lot of effort. However, the most important thing is always that you learn from every experiment that you do, and you keep going." You can learn more about Eddy and his research in our ***Innovation*** 2018 edition.

Research is a never ending process that sometimes presents more questions than answers, but in the end, it has a lifelong impact on the one performing the research.

About the Feature Article Authors

Aditi Kona

Aditi is a rising senior at North Oldham High School. Her passion lies in neuroscience and in her free time, she loves playing piano and dancing.

Yaara Aleissa

Yaara is a senior at duPont Manual and is in extracurriculars that help her enhance her passions of social justice by making an impact on her local community. She enjoys learning about the various cultures that surround her and expand her expertise in the sciences. In her free time you can find her trapped in a good book or a good cooking show.

Shubh Gupta

Shubh is a senior at duPont Manual High School who enjoys learning about math, and honing his public speaking skills. He is involved in several academic teams at school and is passionate about uplifting youth in the local community. In his free time, you can him playing basketball or golf with friends, or filmmaking.

Gokul Achaththekoot

Gokul is a junior at duPont Manual. He has a passion for math and science, public speaking, and volunteering. In his free time, you can find him swimming competetively, playing basketball, or hanging out with friends or family.

Heavy Metal Removal On Industrial Wastewaters

Gia Mendiratta & Taylor Zekoski
duPont Manual High School
11th, Louisville, Kentucky

ABSTRACT

Industries across the nation are incorporating heavy metals into their manufacturing, which creates the challenge of removing heavy metals from wastewater. Cadmium is a heavy metal that causes many health issues. The goal of this experiment was to discover how changes in pH level, loading, and mixing time affect the amount of cadmium particulate removed from treated wastewater in order to provide valuable data that would allow companies to more effectively remove heavy metals. The research sought to prove the hypothesis that optimal cadmium removal would occur at an adsorbent quantity of 125 mg, a mixing time of 20 minutes and a pH between 2-4. Alumina was added to a 15ppm cadmium solution in varying quantities, eventually changing pH and mixing time. The samples were taken to ICP-AES analysis to determine the amount of cadmium removed from each sample. The results for the loadings samples supported the hypothesis as the 125 mg solution absorbed the most cadmium at 92.55%. The results for the mixing time samples supported the hypothesis as the solution with a mixing time of 20 minutes absorbed 92.02% of the cadmium. Since standards were not made with the pH adjusters, the pH results were inconclusive as we had no way of knowing the impact it had on the ICP AES analyses. This experiment concluded that alumina nanoparticles are effective at removing cadmium from wastewater, absorbing over 92% of cadmium under certain conditions.

INTRODUCTION

As the number of industries incorporating heavy metals, such as cadmium, into their manufacturing increases, the costly challenge of heavy metal removal from wastewater becomes direr.

Throughout the country, cadmium is being released into waterways. Cadmium, a bluish-white metal located in zinc ores, has many uses and can be found in several industries: such as batteries and solar cells. However, it is a toxic metal that does not corrode easily. While small amounts of cadmium can be naturally found in foods such as spinach, lettuce, and potatoes, large amounts of cadmium can cause a variety of health issues. Its exposure has been known to cause cancer, as well as target the body's cardiovascular, renal, gastrointestinal, neurological, reproductive, and respiratory systems (Department of Labor logo UNITED STATES DEPARTMENT OF LABOR.). If it accumulates in the blood, mainly done through inhalation, it can cause several symptoms, including high blood pressure, kidney damage, and the destruction of red blood cells (Tripathi and Ranjan, 2015).

Nanoparticles, defined as particles of matter that are between 1 and 100 nm in diameter, are an environmentally friendly and inexpensive, but effective way to remove dangerous heavy metals from the environment. While they have a few harmful environmental effects, they ultimately help the environment as they absorb contaminants and reduce the concentrations of free molecules of pollutants. The heavy metals are attracted to and attach to the nanoparticles because of the different positive and negative charges. Then, the solution is centrifuged, creating a supernatant above the solid residue. The solid residue traps the heavy metals attached to the nanoparticles, removing them from the supernatant above. An ICP machine (Iris Intrepid II) measures the

amount of cadmium particulate in a solution. (Taghavi et al., 2013).

In 2010, an article was published on the "Simultaneous removal of heavy-metal ions in wastewater samples using nano-alumina modified with 2,4-dintriphenylhydrazine"; it tested alumina's effectiveness on the removal of the metal cations Pb(II), Cd(II), Cr(III), Co(II), Ni(II) and Mn(II) from wastewater samples. The article and this study differ as this article is examining the effectiveness at which alumina can remove several differerent types of heavy metal ions, while this experiment is only testing one heavy metal, but is testing how three different factors affect its removal(Afkhami et al., 2010).

This experiment examines the effects of pH level, loading, and mixing time, by using nanoparticles, in this study: alumina, adsorption to determine the amount of cadmium particulate left in treated wastewater. . This experiment will answer the question, "How do changes in pH level, loading, and mixing time affect the amount of cadmium particulate found in treated wastewater?" The results from this experiment will provide valuable data on how changes in pH, loading, and mixing time affect the effectiveness of cadmium removal using nanoparticle adsorption.

The hypothesis of this experiment is as the pH of the solution becomes more acidic, the amount of cadmium particulate left in the solution will decrease; as the loading of the alumina increases, the amount of cadmium particulate left in the solution will decrease; as the mixing time approaches the median, the amount of cadmium particulate left in the solution will decrease. Cadmium Chloride (CdCl2), the molecule in the wastewater solution, is acidic, and alumina is basic. Therefore, they attach and are attracted to each other. If the solution they are in is more acidic, CdCl2 will most likely have a stronger attraction towards the basic alumina. If the loading, amount of alumina, increases, there will be more surface area for the CdCl2 to stick to, so more cadmium will stick to the alumina, and less will be left in the solution. If the mixing time is too short, the solid residue that contains the CdCl2 attached to the alumina will not form below the supernatant, but if it's too long the cadmium particles will detach from the alumina. The results from this experiment will provide valuable data on how different factors affect the efficiency at which cadmium is removed, acting on these insights will allow companies to more efficiently remove heavy metals.

METHODOLOGY

This experiment was conducted in a University of Louisville chemistry laboratory because it required the use of dangerous chemicals. First, a 15 parts per million (ppm) cadmium solution resembling the industrial wastewater containing large amounts of heavy metal cadmium was made. Using mole conversions, the number of milligrams (mg) of cadmium chloride needed for a 400 milliliter (mL), 15 ppm cadmium solution was calculated; 9.8 mg, the calculated amount of cadmium chloride, was measured out using a scoopula to transfer the chemical, a weigh boat to hold the chemical being measured, and a top-loading balance to measure the cadmium chloride. 400mL of deionized water was poured into an Erlenmeyer flask. Then, the 9.8 milligrams of cadmium chloride was poured into the flask with the deionized water, along with a magnetic stir rod. The solution was placed on a magnetic stirrer on low for 15 minutes. Finally, a pH reader measured the pH of the solution, which was recorded. When not in use, the solution remained covered to prevent exposure.

For the loading samples, a pipette was used to transfer 10 mL of the cadmium solution into five plastic centrifuge tubes, used to measure out the 10 mL(each tube will have 10 mL of the cadmium solution). A separate pipette was used to transfer 10 mL of deionized water into five plastic centrifuge tubes (each tube will have 10 mL of deionized water). The measured solutions and water were poured into corresponding jars. Weight boats were used to measure out 25 mg, 50 mg, 75 mg, 100 mg, and 125 mg of alumina nanoparticles. Each of these measurements was poured into one of the jars with the cadmium solution. Each jar and plastic centrifuge tube was labeled with a serial number that will correspond with a sample description in a created chart. Each sample was vortexed for three minutes, so the nanoparticles were sufficiently incorporated. Each sample's pH was measured and recorded. They were then

transferred back to the corresponding plastic centrifuge tubes and were centrifuged for 12 minutes at 4000 revolutions per minute (RPM). Finally, they were set aside in a tube rack.

For the samples that altered in mixing time, there were five plastic centrifuge tubes, each filled with 10 mL of the cadmium solution. Each of the tubes was filled with 75 mg of alumina nanoparticles. They were transferred to corresponding jars, labeled with individual serial numbers, vortexed for three minutes, and measured for pH. After being transferred back into the plastic centrifuge tubes, one of each type of sample was centrifuged for 12 minutes (min), 16 min, 20 min, 24 min, and 28 min at 4000 RPM. The samples were then set aside in a tube rack.

For the samples that altered in pH, there were five plastic centrifuge tubes, each filled with 10 mL of the cadmium solution. Each of the samples was transferred to separate jars and labeled with their individual serial numbers. Each of the jars was filled with 75 mg of alumina nanoparticles and was vortexed for three minutes. Then, pH buffers were added to the solutions until one of each type of sample had one of the following pHs: 4, 6, 8, 10, and 12. All of the samples were centrifuged for 12 min at 4000 RPM and set aside in a tube rack.

The control was a 10mL solution of cadmium chloride without the addition of the alumina nanoparticles. This control sample went through the same conditions as the other sample, except no nanoparticles were added. The solution of the vortex, centrifuged for 12 minutes at 4000 RPM, filtered, and labeled. This sample was used to compare and determine the percentage of cadmium the treated sample removed.

A filter was placed on the syringe used to transfer the supernatant into a new plastic centrifuge tube to be labeled and sent for ICP testing. The ICP testing emits photons that excite atoms and ions with a plasma, causing it to emit radiation at wavelengths characteristic of a particular element, helping to detect how much of the element remains in the sample.

DATA AND RESULTS

The values produced by the alumina loading tests indicate the amounts of alumina that absorbed the most cadmium; the solution with 125 mg of alumina absorbed the most cadmium at 92.549%. The percentage of cadmium absorbed was calculated by subtracting the ICP produced a value of the 15 ppm cadmium solution, 22.94, by the ICP produced value of the solution of 75 mg of alumina in deionized water, 1.707; this equaled 21.233. The table also provides the standard deviation, which shows how close the values in each trial were to each other. All of the loading values had very low standard deviations, except the 25 mg, which means its trial values were more spread out or less consistent compared to the other samples.

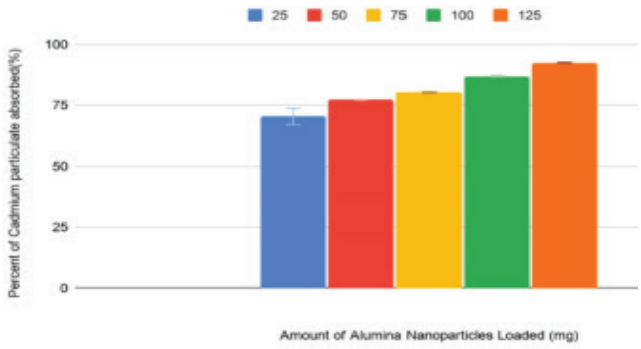


Figure 1
Percent of Cadmium Particulate Absorbed vs. Amount of Alumina Nanoparticles Loaded

Figure 1 compared the percentage of cadmium particulate absorbed to the milligrams of alumina added to each sample. The bar graph shows that as the amount of alumina increases, so does the percentage of cadmium absorbed. The error bars on each sample were fairly small as the majority of the standard errors were less than one. This means that the given mean is an accurate representation of the true population mean. Furthermore, none of the error bars in the graph overlapped, which means that the difference between the samples may be significant.

The t-test values were calculated for the loading amounts closest to each other. The experiment required a two-tailed t-test. The degree of freedom was 4, and the sampling

distribution value was 2.776 at an alpha level of 0.05. When the t-test value is less than the sampling distribution value, the null hypothesis is accepted. The null hypothesis is, "The percentage of cadmium particulate absorbed for both of the loading amounts is equal." When the t-test is greater than the sampling distribution value, the null hypothesis is rejected and the alternative hypothesis is accepted. The alternative hypothesis is, "The percentage of cadmium particulate absorbed for both of the loading amounts is significantly different." All of the t-test values showed significance or rejected the null hypothesis except the test between the 25mg and 50mg trials

The raw data produced by the inductively coupled plasma (ICP) machine and the values calculated using that data with the samples varied in mixing/centrifuge time. The results from the mixing time test indicate that the sample that was in the centrifuge for 20 minutes absorbed the most cadmium at 92.017%. It also illustrates that all of the standard deviations were less than one. Additionally, the error bars on each of these samples were less than 0.5.

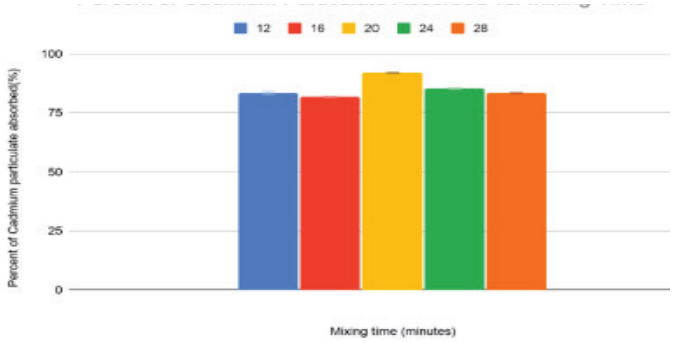


Figure 2
Percent of Cadmium Particulate Absorbed vs. Mixing Time

Figure 2 compared the percentage of cadmium particulate absorbed to the number of minutes each sample spent in the centrifuge. The bar graph shows that the greatest percentage of cadmium absorbed was found at 20 minutes and dropped at the greater and lesser times. Moreover, none of the error bars in the graph overlapped, which means that the difference between the samples may be significant.
The t-test values were calculated for the mixing times closest to each other. The experiment required a two-tailed t-test. The degree of freedom was 4, and the sampling distribution value was 2.776 at an alpha level of 0.05. The null hypothesis is, "The percentage of cadmium particulate absorbed for both of the mixing times is equal." The alternative hypothesis is, "The percentage of cadmium particulate absorbed for both of the mixing times is significantly different." All the tests had t-test values greater than the sampling distribution value, which means the hypothesis was rejected and the alternative hypothesis was accepted.

The results of the pH tests show that the sample that was in the basic solution of a 10 pH absorbed the most cadmium at 94.904 percent. The table also illustrates that all of the standard deviations were less than 0.3. Additionally, the error on each of these samples was less than 0.2.

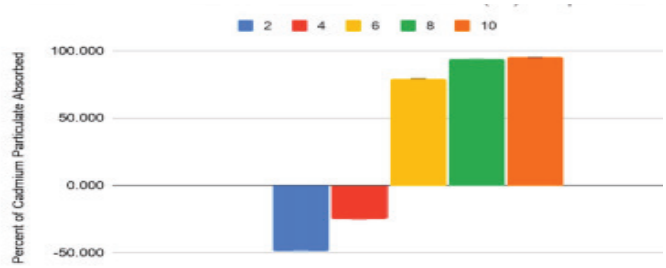


Figure 3
Percent of Cadmium Particulate Absorbed vs. pH Values

Figure 3 compared the percentage of cadmium particulate absorbed to the pH values of each sample's solution. The bar graph shows that the greatest percentage of cadmium absorbed was found at a pH of 10 and lowered as the pH became more acidic. However, these results cannot be used to determine pH's effect on the percent of cadmium absorbed because we did not make our standards with our pH adjusters, so we had no way of knowing the impact it would have on the ICP AES analyses, rendering the pH analysis inconclusive. This explains why some of the percentages of cadmium absorbed were negative.

The results supported this hypothesis. The sample with the smallest amount of alumina at 25 mg absorbed the least amount of cadmium with 70.376%; the percentage of cadmium absorbed increased as the amount of alumina increased. The 125 mg of alumina solution, the solution with the largest amount of alumina, absorbed the most cadmium at 92.549%. The

data showed that larger amounts of alumina absorbed larger amounts of cadmium because as the amount of alumina in a solution increases, there is more surface area for the CdCl2 to adhere to; therefore, more cadmium will adhere to the alumina and less will remain in the solution. All the tests had t-test values greater than the sampling distribution value, which means the results are significant.

The hypothesis for the mixing samples is, "As the mixing time approaches the median, the percentage of cadmium absorbed will increase." The hypothesis was supported by the data as the median sample amount, 20 minutes, absorbed the most cadmium with 92.017%. As the mixing time drifted farther away from the median time of 20 minutes the percentage of cadmium absorbed generally decreased. The median time produced the best results because if the mixing time is too short, the solid residue that contains the CdCl2 attached to the alumina will not form below the supernatant, but if it's too long the cadmium particles will fly off the alumina. All the tests had t-test values greater than the sampling distribution value, which means the results are significant.

The hypothesis for the pH samples is, "As the pH of the solution becomes more acidic, the percentage of cadmium absorbed will increase." This was predicted because cadmium chloride is acidic while alumina is basic. Since opposites attract, if the solution is more acidic, CdCl2 will most likely have a stronger attraction towards the basic alumina. Unfortunately, no conclusion could be drawn from these results because we did not make our standards with our pH adjusters; therefore, we had no way of knowing the impact it would have on the ICP AES analyses, rendering the pH analysis inconclusive. This explains why some of the percentages of cadmium absorbed were negative.

In 2013, a study was conducted on the "Removal of cadmium (II) from simulated wastewater by ion flotation technique". This study examined how different parameters, such as flow rate, frother concentrations, and ionic strength, influence the flotation process and the amount of cadmium being removed. It concluded that the sample with the maximum cadmium removal was 92.1% where ethanol was added at a concentration of 0.4% with a collector-metal ratio of 3:1 and a flow rate of 150mL/min. Since both the cadmium and the collector were effectively removed from the wastewater, the results were deemed promising (Salmani et al., 2013). This study is similar to this experiment because they both focus on the removal of cadmium from wastewater. Furthermore, the results in each experiment showed that cadmium was effectively removed, as the maximum cadmium removal/percentage of cadmium absorbed was above 92%. However, these two studies differ as the 2013 study used the ion flotation technique, and this experiment used the adsorption technique. Both these methods can be considered possible solutions to the cadmium contamination issue.

In conclusion, alumina nanoparticles are effective at removing cadmium from wastewater, as it absorbs over 92% of cadmium under certain conditions. This experiment will provide valuable data on how different factors affect the efficiency at which cadmium is removed, which wastewater producing companies can use to more effectively remove heavy metals.

REFERENCES

Afkhami, A., Saber-Tehrani, M., & Bagheri, H. (2010, May 25). Simultaneous removal of heavy-metal ions in wastewater samples using nano-alumina modified with 2,4-dinitrophenylhydrazine. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0304389410006722?via=ihub

Cadmium.org. (n.d.). Retrieved from https://www.cadmium.org/environment/cadmium-exposure-and-human-health

Department of Labor logo UNITED STATESDEPARTMENT OF LABOR. (n.d.). Retrieved from https://www.osha.gov/cadmium#:~:text=Cadmium and its compounds are,, reproductive, and respiratory systems

Salmani, M. H., Davoodi, M., Ehrampoush, M. H., Ghaneian, M. T., & Fallahzadah, M. H. (2013, February 07). Removal of cadmium (II) from simulated wastewater by ion flotation technique. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3691621/

Taghavi, S. M., Momenpour, M., Azarian, M., Ahmadian, M., Souri, F., Taghavi, S. A., . . . Karchani, M. (2013, November 01). Effects of Nanoparticles on the Environment and Outdoor Workplaces. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4477780/#:~:text=3.5.&text=Thus, even though nanoparticles can,toxic effects of the pollutants.

Tripathi, A., & Ranjan, M. R. (2015). Heavy Metal Removal from Wastewater Using Low Cost Adsorbents. Journal of Bioremediation & Biodegradation, 06(06). doi:10.4172/2155-6199.1000315

Effect of Quaternary Ammonium Compounds on the Growth of Mealworms

Gokul Achaththekoot
duPont Manual High School
10th, Louisville, Kentucky

ABSTRACT

Quaternary Ammonium Compounds (QACs) are antimicrobial agents that are used to eliminate bacteria and are found in products that advertise antimicrobial activity. During the ongoing pandemic, disinfectant agents are being used more frequently to remove unwanted bacteria. To assess the potential health effects on mealworms, 2 QACs (Ammonium Hydroxide (AH) & Benzalkonium Chloride (BC)) were exposed in various concentrations (0.005, 0.01, 0.02 for BC & 1.0, 1.5, and 3.0 for AH) for 15 minutes daily for 2 weeks, and their change in length was recorded. The alternate hypothesis was that as QAC concentration increased, mealworm length would decrease. In addition, 2 positive controls of Lysol (BC- 0.1076%) & Windex Glass Cleaner (AH- 12%) were used to compare the individual QACs effects as these reagents have been studied to show a negative effect on the length in the previous project. Mealworms were used for this project, as they possess similar neurotransmitters to humans, like acetylcholine, which respond to growth status and fumes, but in a much shorter time. The results revealed mealworms exposed to AH have an effect on the length between 90-96% compared to the untreated group, whereas BC groups have an effect on the length between 90-93% compared to the untreated group. The statistical T-tests showed significant differences in all treatments of AH and BC except AH 3.0% & BC 0.02%, and, therefore, the alternate hypothesis was accepted.

INTRODUCTION

Quaternary Ammonium Compounds (QACs or Quats) are antimicrobial agents that are commonly found in many products used for removing bacteria. They have been noted to have adverse effects when exposed to animals in varying concentrations (Bellevue/NYU, 2015). The purpose of this project is to test different concentrations of the main Quaternary Ammonium Compounds against mealworms, with positive controls of actual disinfecting agents. This is based on how QACs provide an effect on human beings because the concentrations are enough to kill bacteria, so their safety concentration could still cause potential animal/human effects. Also, mealworms are used for this experiment as they provide a comparison of how different animals and humans would react on a much faster scale. (Clark, 2016).

QACs are primarily used alone but can be added or combined with other compounds to produce disinfecting agents. They have been added to dishwashing liquids, hand soaps, window cleaners, “all-purpose” cleaners, floor products, baby-care products, disinfectant sprays and wipes, air fresheners, and other cleaning products that advertise antimicrobial activity. Inside liquids, QACs are solid substances. They do not evaporate into the air but do leave behind a residue when used. Different household cleaners contain different combinations of many QACs (Bellevue/NYU, 2015). As these compounds are widely used in common household products, their potential health hazards are often overlooked. In many studies, it has been shown that the QACs can cause mild skin irritations, gastrointestinal issues, asthma, eye and mucous membrane issues upon mists, and are also linked to reproductive issues (Bellevue/NYU, 2015). During the current COVID-19 pandemic, the

products with these QACs are more widely used by people to eliminate unwanted germs. This study will be a continuation of a previous study that looked at how disinfectant wipes played a role in the growth of mealworms, which did have an effect of shortening the mealworms' growth.

The first QAC this experiment will be testing is ammonium hydroxide. Ammonium hydroxide (NH_4OH) appears as a clear, colorless solution. Ammonium hydroxide is very toxic to aquatic life and can cause severe skin burns and eye damage. Additionally, it is considered an environmental hazard and is labeled as corrosive (National Center for Biotechnology Information, 2020). Ammonium hydroxide is found in common household cleaners, used as a food additive, and is also an ingredient in cosmetics for denaturing purposes. Ammonium Hydroxide is found in the common household cleaner: Windex Glass Cleaner, at a concentration of 12% in the solution (SC Johnson Global Safety Assessment & Regulatory Affairs, 2018). The toxicity of this chemical, as mentioned before, can be mitigated by using it in low concentrations. For this experiment, low concentrations will be exposed to mealworms, and masks, eye protection, and gloves will be worn for handling this substance

The second QAC this experiment will be testing is N-alkyl dimethyl benzyl ammonium chloride or benzalkonium chloride. It is a clear, colorless liquid, odorless, and dissolves in water easily. It is used in QACs, as a fungicide, virucide, sanitizer, and disinfectant (United Lab Co., 2015). This experiment will use Lysol Disinfecting Spray as the positive control for benzalkonium chloride, as Lysol Disinfecting Spray contains 0.1076% benzalkonium chloride. It is required to be handled in a fume hood, so it is not to be inhaled, ingested and is not recommended for skin/eye contact (United Lab Co., 2015). However, since this experiment will use very low concentrations of the chemical, it will not require a fume hood while handling. To mitigate the toxicity when handled at a low concentration dissolved in water, masks, gloves, and eye protection will be worn.

This project's test subjects will be mealworms. Mealworms provide a comparison to how animals and humans would react, on a much faster time scale, and they also exhibit human-like reactions to chemical stimuli (Clark, 2016). More so, numerous studies have used mealworms as a basis for testing different compounds or effects of fumes. Mealworms and humans contain similar neurotransmitters in their nervous system that allow them to react to fumes and growth status as humans would do but in a much shorter period, allowing for observation (Collins et al, 2017). Such neurotransmitters, like acetylcholine, play a huge role in the growth and development of animals in both humans and mealworms (Brown & Ingianni, 2006). Additionally, mealworms go through expected life processes, and testing compounds on them allow us to see any discrepancies that could lead to evidence of toxicity.

Currently, work on the effect of QACs has been limited. One recent study tested the exposure of quaternary ammonium disinfectants on mice, and their studies showed that an increase in exposure led to a decrease in the fertility of the mice (Melin et al., 2014). Additionally, another study tested the effect of QACs on acute airways in mice, and the study showed a rise in respiratory pulmonary diseases and a loss of a certain stigma in the respiratory system (Larsen et al., 2012). These studies show the adverse effects of QACs role in producing complications and further disadvantages, but they do not focus on the growth and development of animals exposed to QACs. However, the majority of studies and research done on QACs are more toward the effect of QACs on bacteria, and how to test their antimicrobial activity. This experiment will be focusing on the effect of QACs on the growth and development of mealworms. This project is a continuation of a previous project done that tested the effects of disinfectant wipes on the growth of mealworms. In the previous experiment, the results indicated that disinfectant wipes did have an effect on the growth as higher concentrations of QACs were in the wipe, and a greater time was exposed of the wipe to the mealworm. In this study, the majority of the testing will be comparing the effect of common household cleaners containing the previously mentioned QACs and testing the effect of varying concentrations on mealworms. As per research, no additional work has been done on mealworms specifically to QACs.

In conclusion, the purpose of this project is to assess the effects of varying concentrations

of quaternary ammonium compounds on mealworms, while measuring the effect of additional QAC exposure in common household cleaners on the growth and development of mealworms. This study hypothesizes that the mealworms exposed to QAC compounds that have a higher concentration of the prevalent compounds will have retarded growth compared to the control group and the lower concentration of compounds exposed group. This project will be mentored by Dr. Aruna Vashishta, at the University of Louisville's School of Dentistry in the Oral Immunology and Infectious Disease Department. The mentor will help in acquiring and preparing the dilutions of the compounds used in the project and will also help in designing and interpreting the results. The project will be conducted on the below-ground floor of a residential building in Louisville, Kentucky. Masks, gloves, and other necessary precautions will be worn to conduct this experiment safely.

METHODOLOGY

First, the chemicals benzalkonium chloride and ammonium hydroxide needed for the experiment were obtained through the mentor, Dr. Aruna Vashishta. Then, the chemicals were diluted in their appropriate concentrations by the mentor in their lab, as some chemicals required a fume hood: benzalkonium chloride in 0.005%, 0.01%, and 0.02% and ammonium hydroxide in 6%, 12%, and 18%. Other necessary materials, such as the 250 mealworms, 9 Ziploc® containers for mealworms, a 32-oz Lysol All Purpose Cleaner, a 23-oz Windex Glass Cleaner, and 8 Petri dishes were obtained. Ziploc® containers were weighed on a gram scale and air holes were punched for the mealworms to breathe. Next, the mealworms were placed in 8 containers for positive control, and the testing subjects with 25 mealworms in each. The untreated control group, consisting of 50 mealworms, were placed in their respective Ziploc® container.

After this, the mealworms in their containers were weighed on a gram scale to see the starting overall weight, and the calculation of the Ziploc containers was removed from the equation to derive the mealworms' weight. Next, 10 mealworms were picked out randomly from the testing units and 20 from the untreated control to be measured randomly to get the average mealworm length before the experiment began. This was done to make sure that the weight and length of the mealworms were similar in the experimental group and there was no significant difference in weight and length at the beginning of the experiment. One day was allotted for the mealworm to get accustomed to its environment before testing began, as other experiments involving animal participants required the animal to acquire an understanding of its environment beforehand.

After this allotted time, the desired amount of 10mL of each concentration was placed on each of the Petri dishes, which were respectively placed in the 8 testing containers for 15 minutes once daily. A timer was used to record the 15 minutes, and after the time was up, the Petri dishes were removed from the containers. These 15 minutes were given so that all the mealworms could properly inhale and be exposed to the QACs for the experiment. This process was done every day for two weeks.

Finally, the change in weight and length, was measured for each experimental group as well as the untreated controls. Mealworms were placed in a refrigerator overnight and disposed of the next day. The experiment was repeated for a total of 3 trials to make sure the reproducibility of the experiment was met. The total change in weight, as well as length in each of the experimental groups compared to the untreated control, was recorded on an excel sheet. T-tests and an ANOVA statistical test were conducted to decipher the effect of QACs on mealworm growth.

DATA AND RESULTS

After the experiment was completed, the data and results were recorded in the graphs below. Trial 1 showcases an average increase, whereas trial 2 and trial 3 represent an average decrease in weight by 1.89 g and 2.25 g, respectively. In addition, the percent changes of mass were all significantly different from the control. The controls were put in as 100% change, and each trial's data was accommodated to keep the ratio.

Overall, the trend showcased that mealworm average length decreased in every experimental group. To show significance,

the experimental groups were divided into their categories about their QAC. Windex Glass Cleaner was grouped with Ammonium Hydroxide groups, while Lysol was grouped with Benzalkonium Chloride groups.

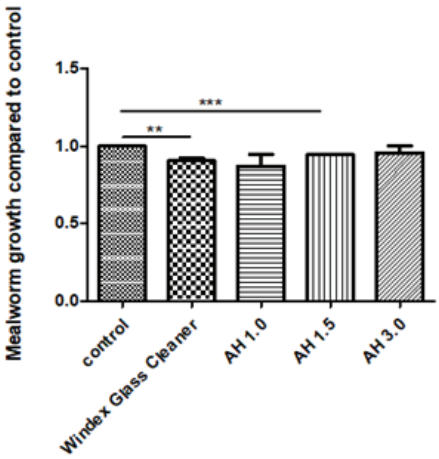


Figure 1:
Ammonium Hydroxide group bar graph representing **Inferential Statistics**

Average mealworm lengths were graphed in Figure 1, and statistical significance was found using a two-tailed t-test comparing each group to the control. In all groups but Ammonium Hydroxide 3.0%, statistical significance was found. Figure 2 was graphed to represent the Lysol & the Benzalkonium Chloride data group.

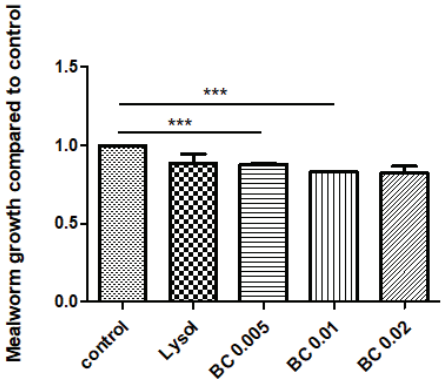


Figure 2:
Benzalkonium Chloride groups bar graph representing Inferential Statistics

In this group, a statistical analysis conducted showed that all groups except Benzalkonium 0.02% were statistically valid. Overall, the main trends include weight overall decreasing and mealworm length decreasing. Also, Lysol groups (0.9171) overall retained a lower percentage in mass when compared to the Windex Glass Cleaner group (0.9200). All t-test significance analyses were taken from 2 degrees of freedom. This trend, similar to the graphs, also shows that the significance in most experimental groups (except AH 3.0 and BC 0.02) was statistically valid. After t-test results and data analyses were complete, results were comprehended.

CONCLUSION

To assess the safety of QAC concentrations used in different disinfectants, this project tested different concentrations of the main QACs used in the disinfectants on mealworms. Different concentrations of QACs are considered toxic and since QACs are found in multiple disinfecting agents, it is vital to understand the effect of these compounds on human health.

Overall, the trends showcased revealed that all of the groups in which any compound (i.e. positive controls and experimental groups) showed reduced growth when compared to the control group. In the Ammonium Hydroxide group, the mealworms showed a 5-10% reduction in length when compared to the control group. However, in Ammonium Hydroxide 3.0% the change in length is not statistically significant as the variation in the experiment was higher. In the Benzalkonium Chloride groups, the mealworms showed a 7-10% reduction in length compared to the control. However, the Benzalkonium Chloride 0.02% (highest concentration) again showed more variation in between experiments and so could not be analyzed.

Additionally, in the Lysol & Benzalkonium Chloride groups, only two trials were reported as there was a heavy loss of mealworms in the third trial. In the case of the Ammonium Hydroxide groups, all 3 trials were considered. Noticing that the Benzalkonium Chloride mealworms grew in a less percentage, it can be concluded that the exposure to a higher toxic chemical (benzalkonium chloride) and the addition of other QACs inside the Lysol spray affected the mealworm's growth. In the previous year's project, the mealworms exposed to Lysol Disinfecting Wipes also grew the least when compared to the other experimental groups. This may be due to the addition of multiple other QACs inside Lysol Disinfecting Wipes,

like ethanol, nonanal, octanol, and lauramine oxide (EWG, 2016). All of these chemicals are known to have multiple side effects, such as general irritation, skin burns, and acute aquatic toxicity. Furthermore, benzalkonium chloride is much more toxic than ammonium hydroxide and is considered corrosive and a major skin irritant (United Lab Co., 2016). For these reasons, benzalkonium chloride showed less growth overall in the mealworms.

Ammonium hydroxide, however, has less effect on the growth of mealworms. Although it is considered dangerous and corrosive, it meets more standards of safety than benzalkonium chloride, as it is in the yellow triangle zone (National Center for Biotechnology Information, 2020). The project showed that as the concentration of ammonium hydroxide increases the variability in the length of mealworms increases and therefore the two lower concentrations showed significance while the highest concentration used in the project did not show significance.

When considering the effects of the overall experiment, mealworm biology may show that the gain in length would be dependent on the stage of the mealworm at the start of the experiment as well as what conditions the mealworm would have been exposed to before starting the experiment. Since these different stages are varied, this may explain why the weight categories were high in variation, from an increase in the first trial to a slight decrease in the second. Mealworms also contract and get shorter as they near their pupa stage, so the growth may have been diminished/minimized to continue this stage (Foss, 2019). Weight gains/losses were not considered for the effect of the QACs as there was no correlation determined. The differences in the toxicity of benzalkonium chloride and ammonium chloride could be due to exposure time and the volatile nature of compounds. Mealworms were only exposed for 15 minutes for 2 weeks, which serve little time to impact overall stage growth. Possible sources of error may have come from a cold temperature in the environment, which may have contributed to diminished growth, and/or the compounds being at varying levels of exposure depending on evaporation rates. At times, the compound may have spilled into the container, which may give for mealworms dying due to ingestion of their food and QAC. These errors were avoided as little as possible, but are essential to consider when looking at the overall explanation. Also, the weight variation shows no trend, so this pattern may have been a possible source of error to see if the mealworms' stage at which they were purchased was consistent. This error is probable as different trials got mealworms from different stages. Trial 2 of benzalkonium chloride groups showed that these heavy losses of mealworms may be due to the condition of the mealworms before experimentation or how they reacted to the environment they were exposed to.

Future research about the effects of QAC could be taken to a level where reproductive success, overall health, or other side effects could be assessed. This may include other animals, as mealworms provide a little outlook into these processes. Higher concentrations, varied concentrations, or a mixture of different QACs may show how overall QACs affect different groups of animals. It is also important to consider testing QAC levels that were found to affect bacterial growth to make sure that these QACs are still able to do their primary function as antimicrobial agents.

REFERENCES

Bellevue/NYU. (2015, September). Quaternary Ammonium Compounds in Cleaning Products: Health and Safety Information for Health Professionals. Retrieved October 9, 2019, from https://med.nyu.edu/pophealth/sites/default/files/pophealth/QACs%20Info%20for%20Physicians_18.pdf

Brown, A. E., & Ingianni, E. (2013, August). No. 43: Mode of Action of Insecticides and Related Pest Control Chemicals for Production Agriculture, Ornamentals, and Turf [PDF]. College Park, MD: University of Maryland. Retrieved September 26, 2020, from http://pesticide.umd.edu/uploads/1/3/5/6/13565116/pil43_modeofaction-agornamturf_2005-2013.pdf

Clark, N. (2016, May 3). Mealworm Behavior Lab. Retrieved September 10, 2019, from https://noahsbiologyblog.wordpress.com/2016/05/01/mealworm-behavior-lab/

Collins, L., Edwards, L., & Chao, S. (2017). Toxicity of Moist Snuff and Impact on Various Stages of Darkling Beetles (Tenebrio Molitor) [PDF]. Fayetteville, NC: Fayetteville State University. Retrieved September 26, 2020, from https://uncw.edu/csurf/explorations/documents/lisacollins.pdf

EWG. (2016, March 14). LYSOL Complete Clean All Purpose Cleaner Spray, Lemon Breeze Cleaner Rating. Retrieved September 6, 2020, from https://www.ewg.org/guides/cleaners/1116-LYSOLCompleteCleanAllPur

poseCleanerSprayLemonBreeze/?formulation=6798
Foss. (2019). Planet and Animal Care: Mealworm. Retrieved September 26, 2020, from https://www.fossweb.com/mealworm
Larsen, S. T., Verder, H., & Nielsen, G. D. (2012, January 18). Airway Effects of Inhaled Quaternary Ammonium Compounds in Mice. Retrieved October 19, 2020, from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1742-7843.2011.00851.x
Lobaugh, E. R. (2019, June 07). Meal Worms: Why Awesome For Research And Experiment. Retrieved September 26, 2020, from https://www.science101.com/meal-worms-research-experiment/
Melin, V. E., Potineni, H., Hunt, P., Griswold, J., Siems, B., Werre, S. R., & Hrubec, T. C. (2014, August 14). Exposure to common quaternary ammonium disinfectants decreases fertility in mice. National Center for Biotechnology Information. Retrieved October 18, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4260154/
Merchel Piovesan Pereira, B., & Tagkopoulos, I. (2019, June 17). Benzalkonium Chlorides: Uses, Regulatory Status, and Microbial Resistance. Retrieved September 27, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6581159/
National Center for Biotechnology Information (2020). PubChem Compound Summary for CID 14923, Ammonium hydroxide. Retrieved September 27, 2020, from https://pubchem.ncbi.nlm.nih.gov/compound/Ammonium-hydroxide.

Medium and Insulator Combination Most Effective to Retain Temperature

Jonah Xu
duPont Manual High School
10th, Louisville, Kentucky

ABSTRACT

Renewable energy has been an emerging field due to the unsustainability of nonrenewable resources such as fossil fuels. Solar energy has become a growing sector of renewable energy, transforming the energy in sunlight to stored heat or electricity. However, solar energy has the issue of intermittence, meaning it is not always readily available. Thermal energy storage (TES) systems can help alleviate this issue. Concentrated solar power (CSP) harnesses solar energy that can heat up a storage medium. This medium is carried to a TES system, which stores the heated medium in an insulator. The purpose of this experiment was to find a combination of medium and insulator that would best retain heat. This led to the hypothesis that a fiberglass insulator and sand medium would be able to retain heat the greatest. In this study, 4 mediums were tested under a CSP device to measure effectiveness to gain heat. Three insulator materials were tested with heated water to measure the ability to retain heat. Sand was found to have the highest mean temperature increase (36.6°C versus 30.8°C oil and 20.0°C water). Fiberglass was found to have the lowest mean temperature decrease (6.4°C versus 8.9°C polyester and 8.0°C styrofoam). All mediums had significantly more temperature increase than the control without CSP while all insulators had significantly less temperature decrease than the control without insulating material. These results indicate that sand and fiberglass are promising materials to use in TES and that CSP is an effective method to increase temperature.

INTRODUCTION

Renewable energy has become an emerging field due to the continual consumption and adverse effects of fossil fuels. The burning of fossil fuels such as coal, natural gas, and oil for energy emits pollutants into the atmosphere. Many of these pollutants increase the amount of solar energy that is absorbed and trapped by the earth's atmosphere, known as the greenhouse effect (Xu & Cui, 2021). The greenhouse effect has disruptive ecological and environmental effects by confining heat in the atmosphere, potentially causing climate change and acid rain. The need for sustainable sources of energy has thus grown, as renewable energy causes little to no environmental pollution and is easily accessed by many people. Solar energy has become a growing sector of renewable energy, transforming the energy in sunlight to stored heat or electricity. Two main methods can be used to generate energy using sunlight, photovoltaic (PV) collector and concentrated solar power (CSP). PV collectors generate electricity directly by using semiconductor materials and displacing particles that are stored or transmitted in the electricity grid. CSP is a method of focusing sunlight using reflectors on a storage medium to generate a large amount of energy in a small area (Gonzalo, Marugán, & Márquez, 2019). Silver mirrors are commonly used in CSP systems, as they reflect near all light that hits the surface of the mirror. However, solar energy has the issue of intermittence, meaning it is not always readily available, especially in the night hours. Additionally, there are increased energy demands of electricity during the night hours and many industrial plants must run on a continuous flow of energy, reducing the potential applications of solar energy. However, thermal energy storage (TES) systems can help alleviate the issue of intermittence. CSP can

heat up a storage medium that is carried to a TES system, which stores the medium with thermal energy that can be tapped at later times. Types of storage mediums include liquids, organic solids, and inorganic solids. TES allows for the storage of transformed solar energy throughout the night, expanding the applications of solar energy (Koçak, Fernandez, & Paksoy, 2020). TES is critical in maintaining the sustainability of solar energy, bridging the temporal gap so stored solar energy can be used even when solar energy is not available (Sarbu & Sebarchievici, 2018). TES also expands the applications of solar energy past electricity production, making direct air heating and water heating possible. Methods to increase the efficiency of TES systems are important, as minimum amounts of energy should be lost when collecting solar energy. The different aspects of TES that are important include the type of medium used and insulation materials used in the storage vessel of the medium.

The thermophysical properties of materials used greatly affect the effectiveness of TES (Alva, Liu, Huang, & Fang, 2017). Thermophysical properties include specific heat, vapor pressure, and conductivity. Other factors like cost and availability are also considered. Materials of particular importance are the storage medium used and the insulator materials for the storage vessel, both of which will be investigated in experimentation.

The experimental question is: "What combination of storage medium and insulator material in solar thermal energy storage will be most effective to retain temperature?". The independent variable is the combination of the storage medium and insulator material. The dependent variable is the amount of time it takes for the insulated storage medium to cool to room temperature, which measures the ability to retain heat. The hypothesis is that the ability to retain heat will be greatest when fiberglass and sand are used. The increase in temperature of different storage mediums (namely water, oil, and sand) along with the temperature decrease over time with 4 types of insulators will quantify the ability to retain heat.

For example, Kumaresan et. al. investigated the use of thermal energy storage for conventional cooking, highlighting an important application of TES. In their paper, fiberglass is named as an insulator for the thermal energy storage unit (Kumaresan et al. 2016). Current experimentation will further test the effectiveness of fiberglass and two other insulators to retain heat.

Another paper collected data from different ways to store water as a medium, providing insight into not only the insulators that can be used but the mode of storage. However, other mediums such as sand were not tested (Slama & Ellasoued, 2018). Current experimentation will continue the discussion of storage methods and limitations of multiple mediums in TES.

METHODOLOGY

Two rounds of experimentation were done for each trial to measure the effectiveness of a medium and insulator combination in thermal energy storage (TES).

Medium preparation

The mediums used in this experiment were water, oil, and sand. For each type of medium, 1.5 cups are placed in an 18.5 x 14 cm shallow glass container, serving as the experimental group. For the control, 0.5 cups of medium was placed into small, circular glass containers to a depth of 1.5 cm. Thermometers were situated in each container such that 5 cm is fully submerged in the medium, ensuring an accurate measurement. After adding the thermometers, a thin transparent covering was applied to each container. The starting temperature of each medium was held constant between the control and experimental groups.

Temperature measurement of the medium in CSP setup

In the concentrated solar power (CSP) setup, different mediums were heated in a CSP device. A CSP device was used in the experiment because this type of device utilizes a medium. It serves as a solar energy collector by heating the medium, which is directly related to the TES systems being investigated. Additionally, concentrating solar power on a medium is representative of real-world solar power collectors, allowing for an increase in

temperature beyond external conditions.

Four flat 18 x 18 cm silver mirrors were used to create a CSP device. Three mirrors were placed upright in a row, with the central mirror angled slightly down. The two adjacent mirrors were also tilted down such that light from the sun was reflected on a small band in front of the central mirror. The fourth mirror was secured to the top of the central mirror, angled further down to reflect sunlight above the bottom mirrors. Tape was used to secure mirrors into position.

The experimental groups were placed inside of the CSP device, such that the concentrated light points directly in the center of the bowl. The control was a medium that was not placed into the CSP device, but still in direct sunlight. The temperature of the control, experimental group, and the environment was recorded in degrees Celsius in 10 minute intervals for 50 minutes. Each time temperature was recorded, the mirrors and position of the CSP device were adjusted to let light fall directly on the medium. In current experimentation, the CSP setup will be effective by using four mirrors to concentrate solar energy onto the medium.

Insulator preparation

The insulators used in this experiment were fiberglass, polyester, and styrofoam. These insulators were placed around each bowl of heated medium in all directions to a thickness of approximately 3 cm. A control group was a bowl without surrounding insulator material.

Measurement of heat loss in medium storage

Insulators are an integral portion of TES, which need insulators to prevent heat loss in storage overnight or in periods without sunlight. Measuring heat loss in an insulator experiment was performed because it tested the effectiveness of the different insulators to retain heat. The insulator with the minimal heat loss was determined using this experimentation method.

In the medium storage experimentation, 0.5 cups of water heated to 50 degrees Celsius was placed into 3 identical circular glass bowls with each insulator with a diameter of 10 cm. The control group was placed into an identical fourth bowl, ensuring that the temperature in each bowl is 50 degrees Celsius. The circular bowls each had a thermometer inside and a thin transparent covering, surrounded by an insulator.

The temperature in degrees Celsius was measured in 10 minute intervals for 50 minutes for the control and experimental groups. Each heat loss experiment was conducted in an indoor setting at a constant temperature of 21.7 degrees Celsius, lower than the outdoor temperature.

This method is ideal because temperature is an easily quantifiable measure of the effectiveness of different insulator materials. This method accounts for differences in medium by making the medium in each insulator water. Each experimentation will be done on 3 types of insulator material and 3 types of medium. Analysis of the extent of heat loss using certain insulators and the increase in temperature of the medium in a CSP device together will determine the effectiveness of a combination of medium and insulator.

DATA AND RESULTS

Temperature of medium in CSP device

Average temperature for water, oil, and sand in the CSP device across all trials (n = 7) was graphed over time, shown in Figure 1. A general trend for all mediums for the average temperature is the general increase of temperature as time progresses through each experiment. Another trend for all mediums is that the rate of average temperature increase decreased over time. Average temperature for water begins to level off towards 50 minutes, but to a lesser extent than oil and sand. Between 0 and 10 minutes, the temperature of water increased on average from 22.2 to 30.0 °C, a difference of 7.8 °C. However, the temperature from 40 to 50 minutes only increased from 40.2 to 42.2 °C (difference of 2.0 °C). Between 0 and 10 minutes, the temperature of the oil increased on average from 24.0 to 39.0 °C, a difference of 15.0 °C. However, the temperature from 40 to 50 minutes only increased from 53.2 to 54.8 °C (difference of 1.6 °C). Between 0 and 10 minutes, the temperature of water increased on average from 23.7 to 41.6 °C, a difference of 17.9 °C. However, the temperature from 40 to 50 minutes

only increased from 58.4 to 60.2 °C (difference of 1.8 °C). Because the temperature increase from 40 to 50 minutes was less than that from 0 to 10 minutes for water, oil, and sand, the rate of temperature increase for each medium decreased over time.

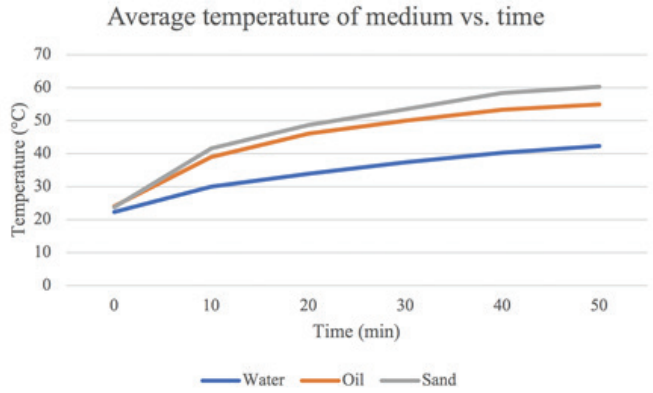


Figure 1

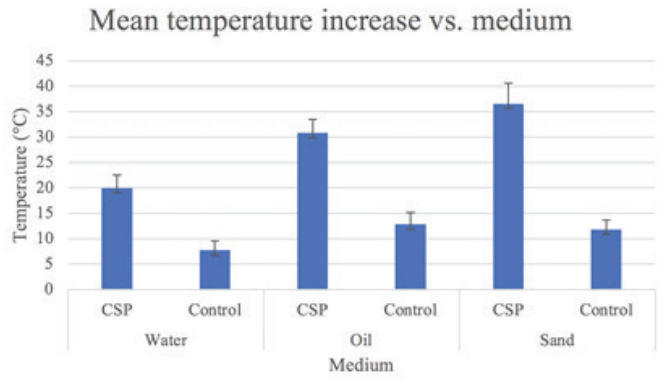


Figure 2

Average temperature increase can be used to quantify the effectiveness of the medium to collect heat derived from solar energy. The average temperature increase along with standard error for each medium is displayed in Figure 2. The positive sign of the mean temperature increase for all mediums rationalizes the general trend of increasing temperature over time for each medium.

The average temperature increase under CSP for water was 20.0 ± 2.53 °C, for oil was 30.8 ± 2.74 °C , and for sand was 36.6 ± 4.00 °C. The average temperature increase of the control for water was 7.8 ± 1.83 °C, for oil was 12.8 ± 2.30 °C , and for sand was 11.9 ± 1.82 °C. Sand had the highest mean temperature increase while water had the lowest. Because the error bars on the graph- shown as one standard error above and below the mean- do not overlap, the mean temperature increase for each medium is significantly different from other mediums. This supports the alternative hypothesis that there will be a greater increase in temperature when using sand over the other two mediums.

The temperature of each medium under a CSP device is significantly different from the control which was not placed in CSP (Pwater = 2.404 × 10-3, Poil = 3.247 × 10-4, Psand = 4.191 × 10-4). The P-value was determined from a student t-test with significance achieved when $P < 0.05$ (threshold value). This statistical test compared the temperature of each medium under CSP and without CSP, supporting that CSP was an appropriate method to significantly increase the temperature of each medium using solar energy.

Measurement of heat loss in medium storage

The average temperature of water in each insulator material across all trials (n = 7) was graphed over time, shown in Figure 3. A general trend for fiberglass, polyester, and styrofoam is that the average temperature of water decreases as time progresses through each experiment.

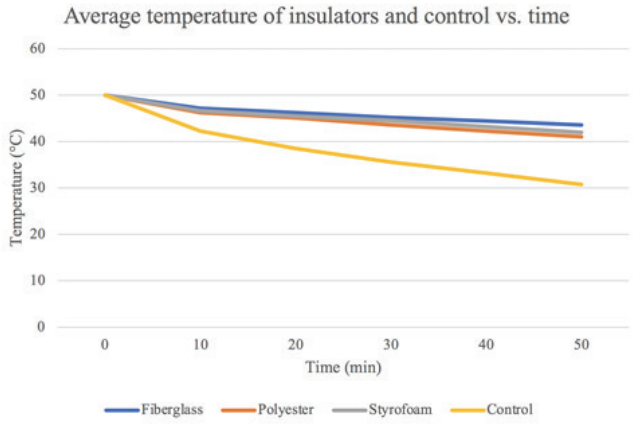


Figure 3

Average temperature decrease can be used to quantify the effectiveness of the insulator to retain heat in the medium. The average temperature decreases along with standard error for each insulator and the control is displayed in Figure 4.

Fiberglass had the lowest mean temperature decrease while styrofoam had the highest. Because the error bars on the graph- shown as one standard error above and below the mean- for all 3 insulator materials overlap, the mean temperature decrease for

each insulator is not significantly different from other insulators. This supports the null hypothesis that there will not be a significant difference in average temperature decrease when using one insulator over the other two insulators.

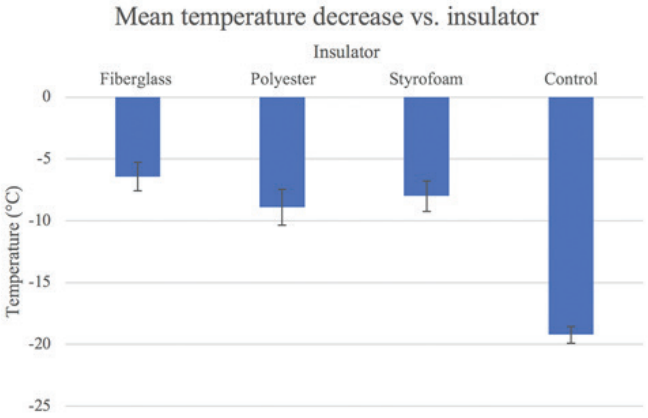


Figure 4

The average temperature of water in each insulator is significantly different from the control which was not placed in an insulator (Pfiberglass = 2.76 × 10-6, Ppolyester = 1.54 × 10-3, Pstyrofoam = 1.91 × 10-5). The P-value was determined from a student t-test with significance achieved when P < 0.05 (threshold value). This statistical test compared the temperature of each medium with and without an insulator. This result is also supported by Figure 3 in that the average temperature for all 3 insulators over time was notably higher than the temperature of the control without an insulator.

CONCLUSION

The need for a dependable source of renewable energy has grown in recent years. Solar energy is a promising sector of renewable energy. Mediums such as water store energy derived from the sun in TES systems. Mediums are transferred to insulators that retain heat so solar energy can be utilized even when the sun is not present. The purpose of this experiment was to find a combination of medium and insulator material that would best store and retain heat, respectively. This ultimately provides beneficial information on the types of materials that can be used in a TES system to best store solar energy.

A general trend seen in the results was each medium in CSP having a higher mean temperature increase over the control, proving that CSP was a successful method in the heating medium. This is rationalized by the positive sign of the mean temperature increase for all mediums. Another trend was each insulator retaining more medium heat than a control without any insulators, rationalized by the control having a large temperature decrease above all other insulator values.

The data confirmed that the sand medium had the highest average temperature increase in a CSP device (16.34 °C compared to 11.66 °C for oil and 6.86 °C for water). The average temperature was shown to be significantly higher than water and oil. This supports the hypothesis that sand would have the highest temperature increase because it had the highest average temperature increase out of all mediums. Data also confirmed that insulators, regardless of material, retained heat significantly more than a control without an insulator (Pfiberglass = 8.53 x 10-5, Ppolyester = 1.43 x 10-3, Pstyrofoam = 2.73 x 10-4, significance value was 0.05). This indicates that insulator materials were successful in retaining heat from the medium. Results show that fiberglass had the lowest mean temperature decrease, meaning it retained heat the best. This supports the hypothesis that fiberglass would best retain medium heat.

For example, Sun et al. used molten salts as a medium in research, showing how other medium types other than water, oil, or sand are used commercially (Sun et al., 2021). Results from this paper support this current experimentation in that the average temperature increase of the medium was more when in a CSP than without. However, this paper had conditions at a much higher temperature suitable for large scale TES. Additionally, Kumaresan et al. investigated heat storage in TES using fiberglass as an insulator, using similar methods in current experimentation. However, their study uses a much thicker layer (0.15m), indicating that thickness of the insulator may be important in retention of medium heat. Physical material properties may explain the results of current experimentation. Specific heat capacity is the amount of heat it takes to warm 1 gram of substance by 1 degree Celsius. A higher specific heat capacity means that it takes more energy to warm the material up compared to other materials with a lesser specific heat capacity. The specific heat capacity of water, oil, and sand are 4182, 1790, and 830 J/kg*C, respectively (Specific heat of some common

substances). This predicts the results from experimentation in that water had the highest specific heat capacity and sand had the lowest. Sand having the lowest specific heat capacity means that it takes the least time to heat up, and that leaving it in CSP for the same amount of time with another medium with a higher specific heat capacity would result in sand having a higher temperature.

Thermal conductivity measures the ability to conduct heat. The thermal conductivity of fiberglass, polyester, and styrofoam are 0.04, 0.05, and 0.033, respectively (Thermal Conductivity of some selected Materials and Gases). A higher thermal conductivity would be expected to allow for more heat to be lost when storing a heated medium. Data support that insulators all have significant retain heat over control. However, experimentation found that there was not a significant difference between the 3 variables. Thermal conductivity would predict for styrofoam to have the highest potential to retain heat, indicating that more experimentation should be done to determine support for this hypothesis.

The varied outdoor environment from one trial to the next in the CSP setup may have affected the results of one trial. However, performing multiple trials may have lessened the effect of this source of error. Alteration in the methodology that would further reduce this potential source of error is to increase the number of trials for both the CSP setup and insulator heat retainment.

Potential further research includes experimenting with different mediums inside of each insulator material along with water. As previously stated, specific heat capacity may influence the ability of the medium to retain heat in an insulator. Doing such an experimentation would provide additional information about each medium in retaining heat, not just gaining heat as was investigated in current experimentation. Additionally, trials can be conducted to extend the time that the medium is in CSP and insulators to more than 50 minutes.

REFERENCES

Alva, G., Liu, L., Huang, X., & Fang, G. (2017). Thermal energy storage materials and systems for solar energy applications. Renewable and Sustainable Energy Reviews, 68, 693-706.

Gonzalo, Alfredo Peinado, et al. "A Review of the Application Performances of Concentrated Solar Power Systems." Applied Energy, vol. 255, 2019, doi:10.1016/j.apenergy.2019.113893.

Koçak, B., Fernandez, A. I., & Paksoy, H. (2020). Review on sensible thermal energy storage for industrial solar applications and sustainability aspects. Solar Energy, 209, 135-169. doi:10.1016/j.solener.2020.08.081

Kumaresan, G., et al. "Performance Assessment of a Solar Domestic Cooking Unit Integrated with Thermal Energy Storage System." Journal of Energy Storage, vol. 6, 2016, pp. 70–79., doi:10.1016/j.est.2016.03.002.

Sarbu, I., & Sebarchievici, C. (2018, January 14). A Comprehensive Review of Thermal Energy Storage. Retrieved September 9, 2020, from https://www.mdpi.com/2071-1050/10/1/191

Slama, R., & Ellasoued, R. (2018, May 11). Conception of Two Wall-Solar Water Heater. Retrieved October 27, 2020, from https://www.omicsonline.org/open-access/conception-of-two-wallsolar-water-heater-2576-1463-1000202-101185.html

Specific heat of some common substances. (n.d.). Retrieved February 20, 2021, from https://www.engineeringtoolbox.com/specific-heat-capacity-d_391.html

Sun, Z., Su, L., Gao, X., Lu, G., Song, X., & Yu, J. (2021, January 25). Influences of impurity Cl− on the thermal performance of solar salt for thermal energy storage. Solar Energy, 216, 90-95. doi:10.1016/j.solener.2020.12.057

Thermal Conductivity of some selected Materials and Gases. (n.d.). Retrieved February 20, 2021, from https://www.engineeringtoolbox.com/thermal-conductivity-d_429.html

Villasmil, Willy, et al. "A Review and Evaluation of Thermal Insulation Materials and Methods for Thermal Energy Storage Systems." Renewable and Sustainable Energy Reviews, vol. 103, 2019, pp. 71–84., doi:10.1016/j.rser.2018.12.040.

Xu, Y., & Cui, G. (2021). Influence of spectral characteristics of the Earth's surface radiation on the greenhouse effect: Principles and mechanisms. Atmospheric Environment, 244. doi:10.1016/j.atmosenv.2020.117908

Determining the Effect of Added Zinc on Plants

Khushi Pola
duPont Manual High School
10th, Louisville, Kentucky

ABSTRACT

The purpose of the study was to determine the effects of added zinc on plant type Vigna unguiculata subsp. unguiculata, on the height of the plants and the chlorophyll content in the leaves. Zinc contamination soil is becoming a prominent issue, as soil is a finite resource and soil pollution impacts crop yields and quality. It was hypothesized that if the amount of added zinc was greater than 200 ppm, plant growth would decrease, and chlorophyll content would decrease. If the amount of added zinc was between 0 and 200, plant growth would increase, and chlorophyll content would increase. To test the hypothesis, forty Vigna unguiculata subsp. unguiculata plants were planted in forty Styrofoam cups, with 100 grams of sterilized organic soil in each cup. The plants were divided into four groups, and each group received 0 ppm, 100 ppm, 200 ppm, or 300 ppm of zinc sulfate fertilizer. Sixteen milliliters of water was given to the plants every day, and height measurements were taken every alternate day. Chlorophyll concentrations were determined through the use of a colorimeter; this was done once at the end of the experiment. Based on the results of the t-tests performed, it can be concluded that there may be an inhibitory impact on chlorophyll content in leaves when zinc is added to the soil ($p<0.05$), but there is no statistically significant impact on adding zinc to the soil on the height of black-eyed peas ($p>0.05$).

INTRODUCTION

Zinc contamination in soil is becoming an ever-relevant issue today. It is vital that the soil present today is taken good care of, since 95% of the foods consumed today are indirectly or directly produced on soils. Soil pollution has also impacted food security; pollution can reduce yields and quality ("Polluting Our Soil," 2018) and human activities such as mining and steel production can increase the amount of zinc released into the environment (Agency for Toxic Substances and Disease Registry, 2015), potentially contaminating soil and water. The purpose of the experiment was to determine the effects of added zinc (in ppm) on plant height (in millimeters) and chlorophyll content (measured using a colorimeter).

Plants are multicellular organisms that use photosynthesis to create their own food. They consist of two main organ systems: the shoot system, which are the parts of the plant that typically grow aboveground, and the root system, which is typically underground, supporting the plant while taking in water and minerals. Plant tissues belong to one of two groups: meristematic and non-meristematic, or permanent, tissues. Meristematic tissue cells are found in regions where the plant is growing (meristems). The following are 3 types of meristems: apical meristems, which allows a plant to increase length of stem and roots, lateral meristems, in which a mature plant can increase its thickness, and intercalary meristems, which allow leaf blades to increase in length. In contrast, non-meristematic tissues have cells that no longer actively divide ("The Plant Body," n.d.).

As mentioned previously, plants use photosynthesis to fulfill their nutritional requirements. Chloroplasts consist of an outer and inner membrane, stroma (fluid found inside the inner membrane where DNA, ribosomes, proteins, and the thylakoid system can be found floating in it), and a thylakoid system (where

chlorophyll is found and where the light reactions of photosynthesis occur).

Chlorophyll is a pigment found in chloroplasts. Two main types of chlorophyll exist, chlorophyll a and b; they are different in their side chains. Chlorophyll a has a side chain of CH_3, and b has a side chain of CHO. Both types are effective photoreceptors because of their structure. Chlorophyll has a porphyrin ring in the middle of the molecule, with a magnesium atom as the central atom. The porphyrin ring serves as the part of chlorophyll that absorbs light energy making it important for photosynthesis (May, 2000).

For plants, zinc is an essential micronutrient. Zinc is a part of many enzymes that regulate growth in plants (Sutradhar, Kaiser, & Rosen, 2016), and it plays a role in immune defense (Cabot et al., 2019). However, both a lack of zinc and excess of zinc is problematic for the plant (Cabot et al., 2019). As Sutradhar, Kaiser, and Rosen (2016) put it, lack of zinc, or zinc deficiency, can be clearly identified in corn plants, with large white stripes on the leaves, and stunted growth being the most obvious symptoms. They also say to make sure of zinc deficiency by analyzing plant tissue. Soil texture, soil temperature and other factors have been linked to zinc deficiency, as well (Sutradhar, Kaiser, & Rosen, 2016). Symptoms of excess zinc, or zinc toxicity, are chlorosis, the yellowing of leaves, and rolling of the leaves, among others (Rout & Das, 2003). Results of zinc toxicity in plants include little or no flower production, stunted root growth, and inhibition of iron translocation in some cases (Rout & Das, 2003).

Due to the fact that both low and high concentrations of zinc proves detrimental to plants (Cabot et al., 2019), zinc homeostasis mechanisms are able to provide just enough zinc to cells at any given time and in any environment. They consist of many proteins that function as transporters and synthesizers of chelators (Assunção, Schat, & Aarts, 2010). Sinclair and Kramer from Ruhr-Universität Bochum in Germany mention in a study analyzing the use of zinc in land plants that zinc homeostasis networks are developed exclusive to the species of the plant, so specific species can "thrive in different soils ranging from extremely Zn-deficient to highly Zn-polluted" (Sinclair & Kramer, 2012).

It is hypothesized that if the amount of added zinc crosses a certain threshold (> 200 ppm), plant growth will decrease, and chlorophyll content will decrease. However, if the amount of added zinc is between 0 and that threshold, plant growth will increase, and chlorophyll content will increase. This hypothesis is a possible answer to the question: How does added zinc affect the growth and chlorophyll content of plants?

Some other studies have been done concerning the effects of high concentrations of zinc and other metals on plants. In one such study, conducted by a group of researchers in Pakistan, vigna radiata plants were hydroponically grown. The plants were grown in pots containing nutrient solutions either with or without zinc. The effect of zinc was observed on plant growth, chlorophyll content, mineral content, and protein content. It was found that zinc concentrations increased the plant growth and chlorophyll content, with 2 µM being the best concentration for the plants. However, it was also observed that phosphorus content decreased as the concentration of zinc changed from 1 µM to 2 µM (Samreen et al., 2017). Sugar beet plants were hydroponically grown in another study, with the various nutrient solutions containing 1.2 (control), 50, 100, and 300 µM zinc, specifically zinc sulfate. It was observed that plants grown with high concentrations of zinc experienced stunted growth, decreased nitrogen, magnesium, manganese, and potassium content, and a damaged root system, among other observations made (Sagardoy et al., 2009).

The experiment to be described will not follow the two studies aforementioned, although the main components were the same: plants and zinc. The plants were grown in sterile soil, not in hydroponics, to help in simulating similar conditions as farmers when planting their own crops.

METHODOLOGY

Black-eyed peas were obtained and planted into a Styrofoam cup. One pea was planted into each cup, which contained 100 grams of potting soil with no added fertilizers. The soil was sterilized by placing it in an oven and baking at a temperature of 180 degrees Fahrenheit for thirty minutes, to remove any

harmful bacteria in the soil. Then, zinc sulfate fertilizer containing 35.5% zinc and 16.5% sulfur was mixed into containers filled with sterilized soil such that the concentration of zinc sulfate fertilizer to soil is 0, 100, 200, and 300 parts per million (ppm), representing the control and the three levels of the independent variable, respectively. Sixteen milliliters of water was given to the plants ten for each level of the independent variable, with forty in total every day after planting of the peas, with height measurements taken a week and three days after planting. The plants were receiving light energy from LED light bulbs placed in lamps for ten hours each day.

For each measurement of the chlorophyll content of each plant, one square centimeter of leaf was cut out of each plant, and placed in five cm^3 of acetone, which was in a glass tube with a screw-on cap. These tubes were refrigerated for five days, as the chlorophyll was extracted from the leaf. The solution was transferred to a cuvette, and using a colorimeter, the absorbance values, measured as absorbance units, at 635 nm were determined. Chlorophyll content was calculated once, with the leaves being cut out three weeks and six days after planting. The experiment lasted approximately one month.

DATA AND RESULTS

The mean plant height of all the dates for the plants in soil containing 200 ppm of zinc were higher than all of the other plants, and the mean plant height for the plants in soil containing 300 ppm were lower than all of the other plants. For example, consider the mean plant height for these groups on December 17th, 2020. For the 200 ppm group, the mean plant height was 116.76 mm, comparative to the mean plant height for the 300 ppm group, which was 80.41 mm. These values were the highest and lowest mean plant heights, respectively. Close to the start of the experiment, it was observed that the mean plant height for the plants in soil containing 0 and 100 ppm of soil were similar, but then the difference increased between the two as the experiment continued.

It can be seen that the trend mentioned previously is shown in this graph: the 200 ppm group had the highest mean plant height, and the 300 ppm group had the lowest mean plant height.

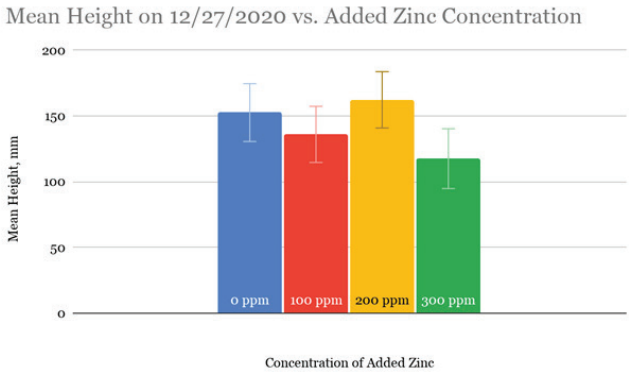


Figure 1: Display of the mean height on the last day of height data collection

It was noticed that the standard error of the mean (SEM) of each value, displayed by the error bars on the graph, were similar to each other, and that visually, the mean heights of the plants in soil containing 0 ppm and 200 ppm are also similar to each other.

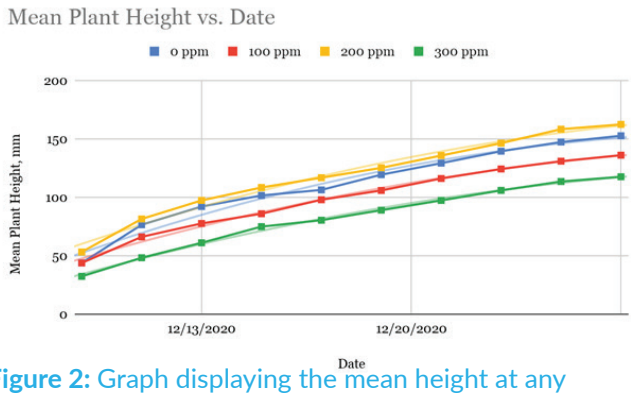


Figure 2: Graph displaying the mean height at any given date, with trend lines

It was made evident that when outfitted with a linear trendline, the slopes were similar to each other. These slopes were 5.47 for the 0 ppm trendline, 4.89 for the 100 ppm, 5.62 for the 200 ppm, and 4.62 for the 300 ppm trendline. Ultimately, a polynomial trendline (with degree 2) was chosen to outfit the graph, because it would simulate growth cycles more efficiently.

It was noticed that the mean absorbance values were similar to each other for the leaf samples from the plants grown in soil containing 200 ppm of zinc and 300 ppm, with the values being 0.0918 and 0.0962 AU, respectively. It was also noted that the mean absorbance value for the leaf samples from the 0 ppm group was much higher than the other absorbance values.

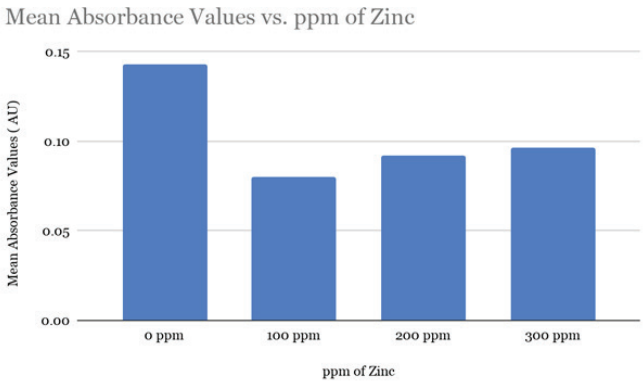


Figure 3: Mean absorvance values vs. ppm of Zinc

It was noticed that the median values were similar to the mean values of their respective data points. For example, the median on December ninth, 2020, for the 0 ppm group was 42.15 cm. In comparison, the mean on that date was 43.87 cm.

It was noted that compared to the medians of the height data, these median values varied more from their respective mean values. It was also noticed that the median values were more similar to each other.

It was noticed that the variance, calculated using the height data from the last day of data collection, for each group was quite high, as well as the standard deviation. It was also noticed that the standard deviation values were similar to each other.

It was noticed that the variance, calculated using the absorbance values, and consequently the standard deviation of the 100 ppm and the 300 ppm groups were the same, and the variance and the standard deviation of the 200 ppm group was similar.

It should be noted that when performing a t-test using the variance value for the 300 ppm group and a different variance value, the t-test value would be greater than 1. The two t-test values that were calculated not using the 300 ppm variance value were both below 1. Comparing the t-values of the height data to the expected value (1.833) resulted in the null hypothesis being accepted, and the research hypothesis unsupported ($p > 0.05$).

It was noticed that all of the t-values were greater than 10, and that the t-values labeled "T-Test: Absorbance: Control vs. 100 ppm" and "T-Test: Absorbance: Control vs. 200 ppm" were similar to each other. Comparing the calculated t-values of the absorbance data to the expected value resulted in the research hypothesis being supported, and the null hypothesis being rejected ($p < 0.05$).

CONCLUSION

Overall, the hypothesis was supported by the findings of two studies: one, concerning hydroponically grown Vigna radiata plants found that zinc concentrations increased the plant growth and chlorophyll content, but that phosphorus content decreased as the concentration of zinc changed from 1 µM to 2 µM (Samreen et al., 2017). In the other study, concerning hydroponically grown sugar beets, observed that plants grown with high concentrations of zinc experienced stunted growth, decreased nitrogen, magnesium, manganese, and potassium content, and a damaged root system, among other observations (Sagardoy et al., 2009).

Throughout the experiment, the plants treated with 200 ppm of added zinc consistently had the highest mean height measurement, and the plants treated with 300 ppm of added zinc consistently had the lowest mean height measurement (see Figure 2). At first glance, this may suggest that 200 ppm of added zinc is the best for Vigna unguiculata subsp. unguiculata, and that 300 ppm of added zinc is the worst for Vigna unguiculata subsp. unguiculata, as predicted by the research hypothesis. However, comparing the t-values for the height measurements to the expected value (1.833) resulted in the null hypothesis being accepted, and the research hypothesis unsupported ($p > 0.05$). Thus, the trends mentioned above concerning the height measurements are not significant enough to be considered actuality.

Another trend observed was in the collection of the absorbance values of the chlorophyll-acetone solutions when light at a wavelength of 635 nm ran through it. The control, despite not having the highest mean height measurement, had the highest mean absorbance value (see Table 2 for quantitative values, and Figure 3 for a visual representation). When comparing the calculated t-values for the chlorophyll content measurements to the expected value, it resulted in the research hypothesis being supported, and the null hypothesis being rejected ($p < 0.05$).

Based on the results of the t-tests, it can be concluded that there may be an inhibitory impact on chlorophyll content in leaves of Vigna unguiculata subsp. unguiculata when zinc is added to the soil, as the mean absorbance values for the experimental groups were all smaller than that of the control group. However, there is no statistically significant impact, whether positive or negative, on adding zinc to the soil on the height of Vigna unguiculata subsp. unguiculata.

Some inconsistencies in data involve the acetone-chlorophyll solution of the control group had an abnormally high absorbance value, compared to the other nine solutions: with that value, which was 0.554, the mean absorbance value was 0.1431 AU; without it, the mean absorbance value was 0.0974, a value that is much closer, difference-wise, to the other means. This may have been due to a technology malfunction in the colorimeter. It was assumed that the first occurrence was merely an outlier, but after this happened multiple times, samples were checked multiple times in the colorimeter to ensure an accurate data point. Thus, the trend mentioned previously concerning the absorbance values may not be fully accurate, and must be revisited. It is recommended that if this experiment were to be repeated, samples are to be checked multiple times.

Another inconsistency was that at least one plant in every group never grew, having no visible stem. When the teardown of the experiment commenced, it was observed that the stem of the plant was growing towards the bottom of the Styrofoam cup; it had curved in on itself, forming a shape similar to the letter "C," with the roots close to the surface of the soil. The cause of this is not believed to be the added zinc, since this was experienced in one of the control plants, as well.

Perhaps there may have been no statistically significant difference in the height of the plants because of the homeostasis mechanisms in plants. Although cases of zinc toxicity are possible, these cases may be possible because there is a higher amount of excess zinc in the soil in those cases than what was used in the experiment. The homeostasis mechanisms in plants in the experiment may have been able to successfully regulate the amount of zinc in the cells of the plant. To further investigate this idea, it is recommended that more zinc should be added to the soil than what was tested, such as 500 ppm or perhaps even 1000 ppm.

For future research, it is recommended that the soil be analyzed for phosphorus, nitrogen, magnesium, and potassium content before and after experimentation, to see if adding zinc affects absorption of nutrients, as seen in other studies (Samreen et al.,, 2017) (Sagardoy et al., 2009). It is also recommended that chlorophyll content be calculated using a spectrophotometer, for more precise measurements, and that the cells of the plants be analyzed for potential damage.

The results of the study may help in determining whether Vigna unguiculata subsp. unguiculata could be used in phytoremediation campaigns. Phytoremediation is a process by which different species of plants are utilized to remove or stabilize contaminants in soil or groundwater. The U.S. Air Force has used poplar trees in groundwater for the purposes of containing trichloroethylene (TCE), and the U.S. Environmental Protection Agency (EPA) has documented that poplar trees "acted as natural pumps to keep toxic herbicides, pesticides, and fertilizers out of the streams and groundwater" ("Phytoremediation," 2014). The specific types of phytoremediation that could be used are phytoaccumulation/phytoextraction, or rhizofiltration. Phytoaccumulation is when the contaminant stays within the stems and leaves, and are good for wastes containing metals. Rhizofiltration does something similar, but is done through hydroponics: contaminated groundwater is used, and as the roots take in contaminants, they are harvested and disposed of ("Phytoremediation, 2014).

REFERENCES

Assunção, A., Schat, H., & Aarts, M. (2010). Regulation of the adaptation to zinc deficiency in plants. Plant Signaling and Behavior, 5(12). doi:10.4161/psb.5.12.13469

Cabot, C., Martos, S., Llugany, M., Gallego, B., Tolrà, R., & Poschenrieder, C. (2019). A Role for Zinc in Plant Defense Against Pathogens and Herbivores. Frontiers in Plant Science. doi:https://doi.org/10.3389/fpls.2019.01171

Chauhan, M. (n.d.). A pilot study on wheat grass juice for its phytochemical, nutritional and therapeutic potential on chronic diseases. [Abstract]. International Journal of Chemical Studies. Retrieved October 18, 2020, from https://www.semanticscholar.org/

paper/A-pilot-study-on-wheat-grass-juice-for-its-and-on-Chauhan/73f2b b7c0b1129b5da66ca284eb2294d956fa9d2?p2df.

Chloroplasts - Structure And Functions: A-Level Biology Revision Notes. (2020, March 27). Retrieved October 18, 2020, from https://alevelbiology.co.uk/notes/chloroplasts-structure-and-functions/

Das, K., Dang, R., Shivananda, T. N., & Sur, P. (2005). Interaction Between Phosphorus and Zinc on the Biomass Yield and Yield Attributes of the Medicinal Plant Stevia (Stevia rebaudiana). The Scientific World JOURNAL, 5, 390–395. https://doi.org/10.1100/tsw.2005.49

Eukaryotic Cells. (n.d.). Retrieved October 18, 2020, from https://www.oercommons.org/courseware/lesson/14952/overview

May, P. (2000). Chlorophyll. Retrieved October 18, 2020, from http://www.chm.bris.ac.uk/motm/chlorophyll/chlorophyll_h.htm

Phytoremediation. (2014). Retrieved January 15, 2021, from http://www.cpeo.org/techtree/ttdescript/phytrem.htm

Polluting our soils is polluting our future. (2018, February 5). Retrieved September 11, 2020, from http://www.fao.org/fao-stories/article/en/c/1126974/

Public Health Statement for Zinc. (2015, January 21). Retrieved September 11, 2020, from https://www.atsdr.cdc.gov/phs/phs.asp?id=300&tid=54

The Plant Body. (n.d.). Retrieved October 18, 2020, from https://www.oercommons.org/courseware/lesson/15091/overview

Rout, G., & Das, P. (2003). Effect of metal toxicity on plant growth and metabolism: I. Zinc. Agronomie, 23(1), 3-11. doi:10.1051/agro:2002073

Sagardoy, R., Morales, F., López-Millán, A., Abadía, A., & Abadía, J. (2009). Effects of zinc toxicity on sugar beet (Beta vulgarisL.) plants grown in hydroponics [Abstract]. Plant Biology, 11(3), 339-350. doi:10.1111/j.1438-8677.2008.00153.x

Samreen, T., Humaira, Shah, H. U., Ullah, S., & Javid, M. (2017). Zinc effect on growth rate, chlorophyll, protein and mineral contents of hydroponically grown mungbeans plant (Vigna radiata) [Abstract]. Arabian Journal of Chemistry, 10. doi:10.1016/j.arabjc.2013.07.005

Sinclair, S. A., &; Krämer, U. (2012). The zinc homeostasis network of land plants. Biochimica Et Biophysica Acta (BBA) - Molecular Cell Research, 1823(9), 1553-1567. doi:10.1016/j.bbamcr.2012.05.016

Stephens, T. J., PhD, McCook, J. P., BS, & Herndon, J. H., Jr., MD. (2015). Pilot Study of Topical Copper Chlorophyllin Complex in Subjects With Facial Acne and Large Pores [Abstract]. Journal of Drugs in Dermatology, 14(6). Retrieved October 18, 2020, from https://jddonline.com/articles/dermatology/S1545961615P0589X

Sutradhar, A. K., Kaiser, D. E., & Rosen, C. J. (2016). Zinc for crop production. Retrieved October 19, 2020, from https://extension.umn.edu/micro-and-secondary-macronutrients/zinc-crop-production

Zinc (Zn) and water. (n.d.). Retrieved October 18, 2020, from https://www.lenntech.com/periodic/water/zinc/zinc-and-water.htm

Zinc. (2003). Retrieved October 18, 2020, from http://chemistry.elmhurst.edu/vchembook/102zinc.html

Developing a Web-Based Convolutional Neural Network for Easy Diagnosis of Pulmonary Tuberculosis

Sai Jeyaprakash
duPont Manual High School
11th, Louisville, Kentucky

ABSTRACT

Every year, Tuberculosis infects 10 million people and kills 1.5 million due to slow and inaccurate means of testing for it, and because Tuberculosis hits hardest in countries lacking medical infrastructure. The purpose of this project was to create a deep learning algorithm that can diagnose Pulmonary Tuberculosis (the most common form of Tuberculosis Disease) via chest x-rays and integrate that model into a web application. This allows for accessibility and speed, as anyone with basic knowledge about computer operation would be able to use this, and because Deep Learning models are extremely fast once trained. The final application allowed users to get on, pick a chest x-ray from their computer through an interactive file window, and upload and get their results with the click of a button. Another Deep Learning model was also integrated to prevent images other than Chest X-rays from being uploaded.

The goal of this project was to achieve an 80% accuracy with the algorithm and a 100% upload success rate with the web application. The final accuracy was 85%, and the upload success rate was 100%, so the engineering goal was met. This project could help towards the effort to reduce tuberculosis infections and help humanity become a step closer to eradicating this disease.

Keywords: Pulmonary Tuberculosis, Deep Learning, Convolutional Neural Network, Web Application, X-ray

INTRODUCTION

Tuberculosis (TB), an infectious disease caused by the Mycobacterium Tuberculosis bacteria, is one of the world's oldest and most widespread diseases. The World Health Organization reports that over 1.4 million people died from TB in 2019 and that it is the leading cause of death from a single infectious agent ("Tuberculosis (TB)," 2020). The situation is also worsening with Multidrug-resistant TB patients increasing by 10% between 2018 and 2019 ("Tuberculosis (TB)," 2020).

One of the major reasons for the difficulty behind controlling the spread of tuberculosis is that there are two kinds of it, Latent TB and TB disease. Although both diseases are caused by the same bacteria, Latent TB is an infection that can't be spread, while TB disease occurs when Latent TB suppresses the immune system and transitions to an infectious form (Lee, 2016). According to one estimate from Houben and Dodd (2016), a quarter of the world's population could have latent TB. Taking into account the large number of people infected with Latent TB, the capability of Latent TB to evolve into TB disease, and the infectiousness of TB disease explains why TB has been extremely difficult to control.

A factor that complicates the problem behind controlling TB disease is the inefficient and inaccurate method of testing it. To test for Pulmonary TB, the most widespread form of TB disease, one has to give two to three samples of their sputum (a mixture of saliva and mucus). These samples have to be stored at a temperature below 32 degrees Celsius and delivered to a microbiology lab where they are screened individually for signs of Mycobacterium Tuberculosis by a specialist (Cudahy & Shenoi, 2016). This process, known as "smear microscopy", has remained largely unchanged for over a hundred years (Nicol, 2010) and is very error-prone, cumbersome, and time-consuming. This process detects only half of all tuberculosis

cases (Mavenyengwa et al., 2016). This means that about half of all Pulmonary TB patients go undetected and spread their disease in the public, which in turn drives more infections. At the turn of the 21st century, researchers began to question the accuracy and efficacy of the smear microscopy process. Gennaro (2000), for example, reported that better methods were needed to diagnose tuberculosis accurately for the purpose of developing and testing potential vaccines. Perkins and Cunningham (2007) emphasized the urgent need for a faster and more accurate method of diagnosis for Pulmonary TB in their study, stating that the HIV epidemic (which is a comorbidity factor for TB) had invalidated the strategy behind using the smear microscopy process (waiting until patients developed severe symptoms) due to the fact that TB was being spread too quickly to keep transmission and mortality under check. Although various methods of testing were proposed and analyzed in many studies like the ones above, none of them even came close to what was required. For example, some of the more modern means of diagnosis, such as the QFT-Gold method and the T SPOT.TB tests have average accuracies of between 70% to 85% (Schluger & Burzynski, 2010). Although these are major improvements over antiquated methods such as smear microscopy, they aren't sufficient for keeping a widespread epidemic under control.

Diagnosis using Deep Learning

In more recent times, several researchers have converged upon the idea of using Deep Learning to diagnose pulmonary tuberculosis. Deep Learning is a method of data analysis where algorithms based on the human brain are fed large datasets (LeCun et al, 2015). Based on the data acquired, the algorithm can then be used to predict outcomes based on incomes. These predictions can be extremely accurate, especially in the field of disease diagnosis. This is due to the fact that a deep learning analysis consists of a highly detailed analysis, which can capture a plethora of details a human may miss (Litjens et al., 2016).

A number of studies have already taken advantage of this powerful technology to create disease-diagnosing algorithms. A 2018 study created an algorithm for diagnosing heart disease and achieved a 96% accuracy rate (Manogaran et al.). This approach has also been applied to diagnosing tuberculosis, and several studies have achieved improved accuracy rates. In the study conducted by Qin et al. (2019), a deep learning model that achieved an accuracy rate of 74% was developed. Another notable example is a study that developed an algorithm with a 79% accuracy at diagnosing tuberculosis based on chest x-rays (Hwang et al., 2019). This shows that deep learning offers the most promising upgrade in diagnosing tuberculosis.

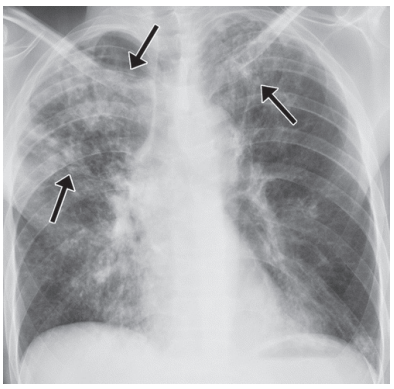


Figure 1: Examples of high contrast pixels occurring in chest x-rays due to PTB

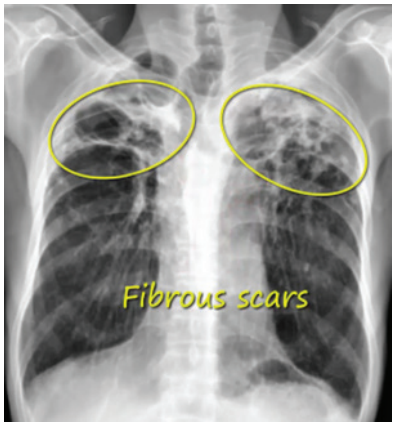


Figure 2: Fibrous scarring in the chest x-ray of a PTB patient

Deep Learning algorithms take advantage of the fact that TB causes inflammation in the lungs, and this causes high contrast pixels to appear in the chest x-ray, such as in Figure 1. When an algorithm detects enough high contrast pixels to pass a pre-set threshold, the x-ray is flagged as PTB positive. Furthermore, especially in severe cases of PTB, fibrous scarring may occur in certain regions of the lungs. Fibrous scarring of the lungs is essentially the formation of scar tissue to heal damage, except that the scar tissue doesn't fade away. This can cause various complications such as blood clots, lung collapse, lung infections, and

even death if untreated. Fibrous scarring is highly visible, even to the human eye, as shown in Figure 2.

Gaps in current technology

While deep learning algorithms are heavily improving the accuracy and speed of diagnosing Pulmonary Tuberculosis, also known as PTB, they are still not accessible. According to a 2019 study, developed deep learning models are not being enabled for widespread accessibility and use, which severely impacts their usability (Qin et al., 2019). Furthermore, most developed algorithms don't use optimization techniques such as removing background noise and other unnecessary details, which could impact the performance of algorithms (Rachna & Swamy, 2013). The aim of this project is to fill these gaps and develop a Convolutional Neural Network that can analyze chest x-rays for signs of Pulmonary Tuberculosis, and integrate it into a web application (a type of deep learning algorithm) that is accessible and highly accurate. The model will be made more accurate than others by using multiple optimization techniques (removing background noise, refining images, etc.). The algorithm will also be more accessible because it will be integrated within a web application, which means that anyone would be able to upload a chest x-ray and get highly accurate results within seconds. This would enable mass-testing campaigns to be executed more easily and inexpensively.

METHODOLOGY

In order to achieve the aims of this project, two deep learning models were developed. The first one was to diagnose Pulmonary Tuberculosis in chest x-rays, and the second one was to prevent images other than chest x-rays from being uploaded to the application, as this may interfere with the application's performance. These models were then integrated into the web application in order to make a fully functioning application. The model for diagnosing pulmonary tuberculosis was trained on a dataset from the National Library of Medicine website, and all of it was thoroughly de-identified.

This project was conducted based on a Non-Experimental Engineering Design, and the Engineering method for AI Software Development was used. This method was custom-created and involved first developing the AI component and thoroughly testing it, and then creating any other components (in this case, the web application). Then, all the components were integrated in a seamless manner. This method was partly inspired by Hwang et al.'s study, which dealt with using deep learning to diagnose heart disease via cardiograms. The prerequisite to completing this web application was to have an organized file structure and sort each image into separate folders based on whether it was positive or negative for Pulmonary Tuberculosis. The same needed to be done for the dataset used to train the second model to recognize chest x-rays apart from other images.

The first and foremost step after meeting the prerequisites was to import all the required python libraries using the following import statements. After this, both Machine Learning models needed to be created and saved. This was done in two separate code files (one for each model), and the primary libraries used were TensorFlow and Keras. During the training of the CNNs, several steps had to be confirmed as done in order to ensure proper model performance.

The first step was the Convolution operation, which is used to remove unnecessary features from the image and make it simpler, converting the image into what is known as a Feature Map. The next step was to apply the ReLU layer, which breaks up linearity in the image caused by applying the convolution operation. The third step was pooling, which is done to ensure that important features of the image are detected no matter where they are located in the image. The fourth step was Flattening, which transforms the pooled feature map into a column of pixels, which makes it easier for the deep learning model to be trained on it. The final step was to deploy a number of hidden layers onto the model and pass each pixel into the hidden layer neurons, which can be compared to nodes that are in charge of looking at specific features of the image and activating if it finds the feature. The model trains itself over and over on every single image in the dataset to figure out which neurons activating corresponds to which final output. Once this was done, the model was finished. In Figure 3 is an example of a CNN that is intended

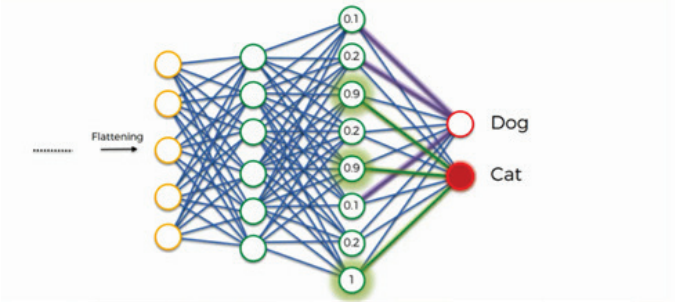


Figure 3: A sample covolutional neural network

to classify images of Cats and Dogs.

The initial column is the flattened feature map that is fed into the model. The layers in between are the hidden layers, and each green circle is a neuron. This image shows an example of how each neuron activates and shows the probability of it correctly identifying the feature in an image. Depending on the value of the probabilities and which of the neurons activated, the model makes a decision about which class the image belongs to. Once the models are created and trained on their respective datasets, the .save() method can be used so that the model doesn't have to be trained each time the website is used. Using the .save() method was a major turning point/breakthrough in this study. Prior to this, the model had to be fully trained every time the website was loaded, which took up a tremendous amount of time. The study done by Litjens et al. mentioned the .save() method as very helpful for this purpose.

After the models are trained and saved, work on the web application was started. The first page that needs to be created is the landing page, and this can be almost anything. It only needs to have clear directions to get to the upload page. This file was titled index.html, and the Flask route to it is simply "/" since it is the default page. The accepted method for this should only be GET since POST is not required here. The next page on the web application was the upload page, and the Flask route can be anything as long as it's clear that this is the upload page. The allowed methods on this page were "GET" and "POST", as both are required on this page. The page had to be able to re-train the model using the saved files from previously, and the training here took less than 5 seconds. Then, the user needs to be able to upload an image using an interactive file window, and the page needs to first analyze if it's a chest x-ray or not. If it isn't, then the user needs to be redirected to a page where it notifies them of a possible mistake and lets them override the action if they think it's a mistake. If the upload is a chest x-ray, then the user needs to be able to see whether or not the chest x-ray is positive or not for tuberculosis. This was done by designating class indices in the code and then using those class indices to figure out what the meaning of the code output is. The final set of code files is given in the appendix of this paper.

For each image upload, the actual accuracy according to the dataset was recorded. Then, five trials were conducted for each image. Each trial involved uploading the image to the web application and recording the model's prediction and upload status (whether or not the image was uploaded correctly). And then, an overall prediction was selected depending on which prediction the model made the most for each image. An overall upload status was also chosen by a similar process. This was repeated for each of the 132 chest x-rays in the test set, 66 of which were Pulmonary Tuberculosis positive and another 66 which were Pulmonary Tuberculosis negative.

DATA AND RESULTS

After applying the data collection method described above for each of the images in the test set, the final prediction accuracy for the model was shown to be 85%, along with a 100% upload success rate.

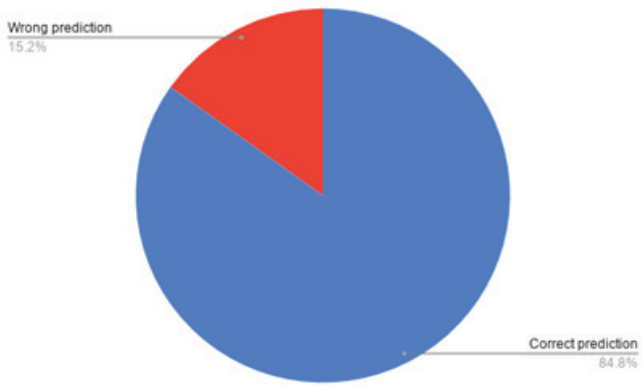


Figure 4: A pie chart compiled from the results that depict the number of correct and wrong predictions

A prominent trend that was found in the data was that the model, although not always correct in its prediction, was always consistent.

If the model's prediction for an image was negative, then the rest of the four predictions were bound to be negative. The same trend applied to positive predictions. Another unexpected result was the 100% upload success rate, due to the fact that no optimization frameworks or code processes were applied to make the uploads consistently successful. This can be attributed to the method, which was the Engineering method for AI software development. This method placed greater emphasis on developing the AI component and other components separately (the Web Application component in this case), and then spending almost as much time on ensuring they work together seamlessly. The main reason for the 100% upload success rate was most likely due to the fact that a lot of time was dedicated to making sure they work together, even if no file upload optimizations were used.

Another prominent trend in the data was that the model had more failures at analyzing Pulmonary Tuberculosis accurately on chest x-rays from patients that lived in urban areas, as shown in Figure 5.

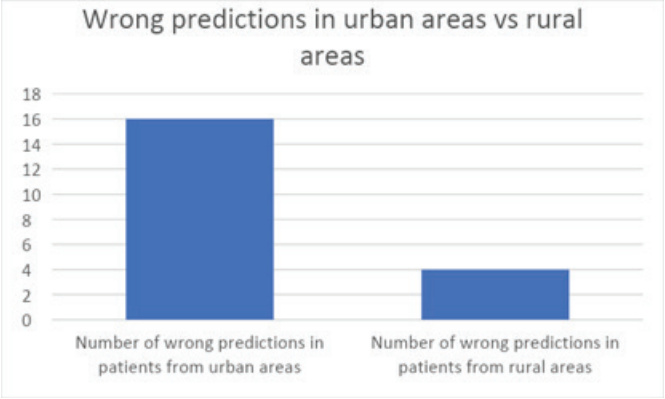


Figure 5: A bar graph comparing prediction failures from urban vs rural patients

According to Fang et al., this can be attributed to the fact that disease spreads more in urban settings. This allows more cases where there are variants of the most widespread strain of Tuberculosis to spread. Since the model isn't trained to accommodate different strains of the disease, this may be affecting the model's accuracy. The reason that this isn't the case in rural areas is due to the fact that rural areas, in general, have a low population density, so the disease spreads less and in turn, mutates less. Although this trend should apply everywhere on the planet, it should be mentioned that this dataset was acquired from China. Due to the high population density of China, this trend may not be as obvious elsewhere in the world when comparing urban and rural results.

Overall, there are many factors that impact the model's accuracy. However, the model seems to work well overall. As for the second model that is supposed to detect images uploaded that are not chest x-rays, that model achieved a 98% accuracy. A lot of time wasn't spent on this model as it wasn't vital to the project's completion. The high accuracy can be attributed to the fact that chest x-rays are very different from most images. The model could take advantage of this to be able to differentiate between images easily and efficiently.

CONCLUSION

After executing the method, new conclusions and understandings could be drawn using the results generated from the data collection process and these results could be compared with those of previous studies. The researchers in these previous studies have been able to achieve accuracies that are above 70% while trying to create deep learning algorithms to diagnose Pulmonary Tuberculosis. However, the fact that image refinement and training optimization techniques were used in this project enables it to achieve higher accuracies. This is because Hwang et al.'s study ignored the use of these techniques, and the rest of their algorithm was almost the same in terms of architecture, quality of data, and size of the dataset. Based on this, it can be said that these image refinement and training process optimization techniques have a large impact on overall algorithm accuracy. The accuracy of the algorithm developed in this project was 85%, which both meets the engineering goal and exceeds the accuracy of previous algorithms made for this purpose by a large margin.

Furthermore, the successful implication of web application integration makes this form of testing a lot more practical and accessible. The web application was able to consistently accept uploads while being able to differentiate between images that were chest x-rays and those that weren't. The web application achieved a 100% upload success rate, which was part of the

engineering goal.

However, a major problem that has been noticed is that this model may be less effective at detecting tuberculosis in urban areas, as mentioned previously. Although this problem can easily be solved by simply analyzing more chest x-rays in the training set, it is still undeniable that this may cause problems. Further research must be conducted regarding how to mitigate this; however, this is perhaps the most unique problem that has come out of this project.

Limitations

There were several notable limitations that caused problems in this project. One of the biggest limitations was the lack of chest x-rays. Although the model was trained on approximately 500 images, many models are trained with tens of thousands of images. Another notable limitation was that there was no variation in the quality of the chest x-rays. Although it is recommended to train a model on high-quality data, simulating real-life conditions might have been more practical and useful for this project. The last major limitation was a lack of data from several countries that have high cases of Tuberculosis. Having data from more regions around the world would increase the model's overall performance and could have solved the problem with urban-rural discrepancies in the model's analysis.

Implications

These results show that a future priority needs to be training the model on more chest x-rays and investigating the effects of this. A second factor that needs to be investigated is how increasing the number of epochs each time a model is trained could affect the accuracy of the model. Finally, as aforementioned, the urban-rural difference inaccuracy is the biggest problem with this application currently. This difference in accuracy must be investigated further and, in order to eliminate this accuracy difference, more chest x-rays from urban environments need to be added to the training dataset of the model. This would allow for a more accurate model while also reducing the accuracy difference between urban and rural environments. Results from these investigations might make this approach to testing Pulmonary Tuberculosis even more accurate and efficient. Overall, this model is highly successful at both analysis and accessibility and takes humanity one step closer to defeating Tuberculosis.

REFERENCES

Basic TB Facts. (2016, March 20). Retrieved October 24, 2020, from https://www.cdc.gov/tb/topic/basics/default.htm

Campbell, I. A., & Bah-Sow, O. (2006). Pulmonary tuberculosis: Diagnosis and treatment.
British Medical Journal, 332. doi:10.1136/bmj.332.7551.1194

Cudahy, P., & Shenoi, S. V. (2016). Diagnostics for pulmonary tuberculosis. Postgraduate
medical journal, 92(1086), 187–193. https://doi.org/10.1136/postgradmedj-2015-133278

Elhoseny, M., & Shankar, K. (2019). Optimal bilateral filter and Convolutional Neural Network
based denoising method of medical image measurements. Measurement, 143, 125-135. doi:10.1016/j.measurement.2019.04.072

Fang, H., Chen, J., & Rizzo, J. (2009). Explaining Urban-Rural Health Disparities in China.
Medical Care, 47(12), 1209-1216. Retrieved March 12, 2021, from http://pac.lfpl.org:2067/stable/27798386

Gennaro, M. (2000). Immunologic Diagnosis of Tuberculosis. Clinical Infectious Diseases, 30,
S243-S246. Retrieved November 24, 2020, from http://www.jstor.org/stable/4482230

Goodfellow, I., Bengio, Y., & Courville, A. (2017). Deep learning.
Cambridge, Massachusetts:
MIT Press.

Grinberg, M. (2018). Flask web development: Developing web applications with Python.
Sebastopol, California: O'Reilly.

Houben, R. M., & Dodd, P. J. (2016). The Global Burden of Latent Tuberculosis Infection: A
Re-estimation Using Mathematical Modelling. PLoS medicine, 13(10), e1002152. https://doi.org/10.1371/journal.pmed.1002152

Hwang, E. J., Park, S., Jin, K. N., Kim, J. I., Choi, S. Y., Lee, J. H., ... & Deep Learning-Based
Automatic Detection Algorithm Development and Evaluation Group Kim Dong Hyeon Woo Woo Choi Choi Hwang In Pyung Song Yong Sub Lim Lim Kim Kim Wi Jae Yeon Oh Su Suk Kang Mi-Jin. (2019). Development and validation of a deep learning–based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. Clinical Infectious Diseases, 69(5), 739-747.

Islam, M. T., Aowal, M. A., Minhaz, A. T., & Ashraf, K. (2017). Abnormality detection and
localization in chest x-rays using deep convolutional neural networks.

arXiv preprint arXiv:1705.09850.
Jo, T., Nho, K., & Saykin, A. J. (2019). Deep Learning in Alzheimer's Disease: Diagnostic
Classification and Prognostic Prediction Using Neuroimaging Data. Frontiers in Aging Neuroscience, 11. doi:10.3389/fnagi.2019.00220
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.
Lee S. H. (2016). Tuberculosis Infection and Latent Tuberculosis. Tuberculosis and respiratory
diseases, 79(4), 201–206. https://doi.org/10.4046/trd.2016.79.4.201
Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., ... & Van Der
Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports, 6, 26286.
Millard, J., Ugarte-Gil, C., & Moore, D. (2015). Multidrug resistant tuberculosis. BMJ: British
Medical Journal, 350. Retrieved October 25, 2020, from https://www.jstor.org/stable/26520021
Mishra, S., Dash, A., & Jena, L. (2020). Use of Deep Learning for Disease Detection and
Diagnosis. Bio-inspired Neurocomputing Studies in Computational Intelligence, 181-201. doi:10.1007/978-981-15-5495-7_10
Nardell, E. A. (2015). Transmission and Institutional Infection Control of Tuberculosis. Cold
Spring Harbor Perspectives in Medicine, 6(2). doi:10.1101/cshperspect.a018192
Reichler, M. R., Khan, A., Sterling, T. R., Zhao, H., Moran, J., McAuley, J., . . . Mangura, B.
(2018). Risk and Timing of Tuberculosis Among Close Contacts of Persons with Infectious Tuberculosis. The Journal of Infectious Diseases. doi:10.1093/infdis/jiy265
Sathitratanacheewin, S., Sunanta, P., & Pongpirul, K. (2020). Deep learning for automated
classification of tuberculosis-related chest X-Ray: Dataset distribution shift limits diagnostic performance generalizability. Heliyon, 6(8) doi:10.1016/j.heliyon.2020.e04614
Tuberculosis. (n.d.). Retrieved October 24, 2020, from https://www.afro.who.int/news/tuberculosis
Welcome to Flask¶. (n.d.). Retrieved October 26, 2020, from https://flask.palletsprojects.com/en/1.1.x/
Yan Guo, Jin Zhang, Chengxin Yin, Xiaonan Hu, Yu Zou, Zhipeng Xue, Wei Wang, "Plant
Disease Identification Based on Deep Learning Algorithms in Smart Farming", Discrete Dynamics in Nature and Society, vol. 2020, Article ID 2479172, 11 pages, 2020. https://doi.org/10.1155/2020/2479172
Zak M., Krzyżak A. (2020) Classification of Lung Diseases Using Deep Learning Models. In:
Krzhizhanovskaya V. et al. (eds) Computational Science – ICCS 2020. ICCS 2020. Lecture Notes in Computer Science, vol 12139. Springer, Cham. https://doi.org/10.1007/978-3-030-50420-5_47
Zaman, K. (2010). Tuberculosis: A Global Health Problem. Journal of Health, Population and
Nutrition, 28(2). doi:10.3329/jhpn.v28i2.4879

Building a Cardboard Prosthetic Hand

Saif Gopang & Salman Shahzad
duPont Manual High School
10th, Louisville, Kentucky

ABSTRACT

Nearly 10% of 2 million amputees in the United States have wrist or hand amputations, driving the need for cheaper prosthetics to provide affordability for low-income amputees and other populations. Thus, the purpose of the engineering project is to construct and operate an EMG-powered prosthetic hand that aims to use cheap building materials to go in the right direction of filling the needs and issues of existing prosthetics. The engineering goal is that all of the cardboard hand's tested attributes will remain statistically insignificant to non-cardboard prosthetic hands in replicating finger flexion. The design process starts with the construction of the cardboard hand, by individually constructing each digit, the palm, and forearm. Thus, the electrical components of the Arduino board, breadboard, servo motors, and wires are assembled to the forearm, remaining accessible to the programming computer. The code developed consisted of rotating the servo motor based on the values collected by the muscle sensor. Prior to the design and coding development, the data collected consists of the main attributes of the prosthetic hand such as speed, strength, range of motion, weight, and cost will be collected and compared to existing non-cardboard prosthetics. Upon the completion of experimentation, statistical analysis shows a significant difference in degrees per second and strength, while showing an indifference between the range of motion, thus not satisfying the engineering goal. In conclusion, the use of cardboard, and cheap materials lead to the liable speed and strength, while improving the range of motion, weight, and cost.

INTRODUCTION

Amputation is the removal of a body part due to trauma, disease, or surgical complications. Amputees are assigned a prosthesis - an artificial body segment or limb that acts as a rehabilitation device and compensates for the loss of a limb's function. The three main types of prosthetics include: cosmetic, body-powered, or myoelectrical prosthetics, which all take nearly 2 to 6 weeks to manufacture, customize, and fit. The overall size, autonomy with working muscles, durability, and haptic feedback are all major factors in determining if a prosthetic hand is appropriate for patient use. The main way hands are still able to operate after an amputation is through a myoelectric prosthetic. These prosthetics usually range from $30,000 upwards towards $100,000 which deters many towards cheaper options (Lam, n.d.). The steep price tag is a result of the arm utilizing movements from existing muscles to control the prosthetic.

The control of a prosthetic hand comes from the use of electrodes. Muscle sensors are used externally on the skin to detect muscular movements through sensing their electrical activity, also known as electromyography (EMG). EMG muscle sensors specifically measure a muscle's electrical activity stimulated from nerve cells during a contraction. Also, when the muscle is at rest, no voltage will be outputted from the sensor (Electromyography, n.d.). These sensors are generally used to aid in diagnosis of muscle or nerve conditions, used for researching kinesiology, and prosthetics. However, the EMG signal is vulnerable to noise and distortion, which are electrical signals that may interfere with the desired EMG signal, or a signal which is altered due to inconsistencies in methodology and muscle activation. Therefore, if the EMG uses non-invasive surface electrodes, the EMG signal may be susceptible to inconsistency as it measures a broader variety of muscle fibers under the available skin in which the electrical

signaling is more random (Raez, Hussain, & Mohd-Yasin, 2006). These sensors cost less than $100 while being able to replicate similar sensitivity and detection capability to the myoelectric prosthesis. The advantages of the use of EMGs include simplistic use, a comfortable fit, and being able to control motors through the user's existing muscles.

In order to function, the prosthetic needs to be completely cohesive in regards to construction, motors, and the sensors. The accurate control of these signals requires the use of expensive, and often, large external hardware. However, Arduino fills this void seeing as it independently functions, is portable, and still maintains accuracy. Also, Arduino allows the hand for ease of compatibility and low cost, aligning with the objectives of this project (D'Ausilio, 2011). Moreover, Arduino can be programmed in many different languages including Java which will be used for this hand.

Servo motors are devices with output shafts, these shafts undergo angular changes dependent on the coding signal. Stepper motors, in contrast, contain high pole counts which result in high torque without the need for an encoder.

The transfer of data through the functioning of the hand starts with myoelectric sensors receiving electrical signals during flexion of the fingers. The Arduino board then computes this signal into a value which indicates the force of the motors which results in the flexion of the fingers. (Esposito et al., 2018). The Arduino can compute the input of the muscle sensor due to the programming embedded to the board. The programming of the Arduino board outputs signaling into the servo motors, which is dependent on the strength of the electrical activity sensed by the muscle sensors. However, to prevent the hand from being too weak or too strong, parameters will be set into place to minimize malfunctions or damage to the hand. Building techniques often revolve around the choice of exterior material. Furthermore, 3D printed plastic prosthetic hands are emerging in the manufacturing scene, which have low material usage, cost, ease of use, and customizability. However, these prosthetics face the problems of low durability, insufficient materials, possible unregulated temperature leading to cracks, and the inability to create larger bunches. Materials such as cardboard also contain construction problems such as deformation, and weather vulnerability. However, cardboard has advantages such as being biodegradable, sustaining durability over time, lower energy consumption, and environmental impact, making it feasible for the purpose of this project.

According to Johns Hopkins Medical School, nearly two million people in the United States live with amputations, and 10% have wrist or hand amputations ("Amputation," n.d.). However, only 37% to 56% of upper limb amputees use prosthetic devices (Raichle et al., 2008). Eighty percent of amputees come from low-income countries where 1% of the population has access to rehabilitation services, making it unattainable for a large portion of amputees (Chan & Zoellick, 2011). Prosthetics also require frequent maintenance with repairs and adjustments.

These problems extend to the potential amputees who can't afford cosmetic or functional prosthetics to face a low quality of life, worse mental health, lack of mobility, and loss of functionality. The potential change that can be made to existing prosthetics is to lower the cost of manufacturing. This allows for cheaper prosthetics and more employment of highly skilled professionals, thus increasing the production rate.

The goal is to operate a cardboard exterior hand controlled by EMG that is lightweight, low cost, and easy to use. Furthermore, this experiment aims to reduce wait time and create a more efficient device. This cardboard prosthetic hand should be able to use motors to contract and expand four fingers and the thumb simultaneously.

This project aspires to build a prosthetic hand that will consistently complete and is successful in completing day-to-day tasks. Success using this method can be seen with the Southampton Hand Assessment Procedure (SHAP) that assesses the ability of the prosthetic to grip and move through the completion of tasks (Merrett, n.d.). The range of motion of the fingers will be recorded to find the wanted tendon tightness that will be compared to already made prosthetics.

The first previously published work related to this project is by M.C. Carrozza, B. Massa, S.,

Micera, etc. Their prosthetic hand replicated the human hand through functionality and feeling. A second related project includes the work from Corinne Dally, Daniel Johnson, Moriah Canon, Sarah Ritter, and Khanjan Mehta. The goal of the work was to test the prosthesis with daily tasks. A third related publish is from Kathleen Talbot, where similar mechanisms in the use of muscle sensors, Arduino Uno microprocessor board, step motors, and a muscle sensor were used to operate their prosthetic. The research of these scientists relates to this project as the tests used to measure the hands' capabilities are similar to this project's procedural tests to evaluate if the built prosthetic hand reaches the engineering goal. The current experiment fits into the gaps of current research as the previous project tests the percentage of succession towards common tasks using a 3D-modeled hand, while this project is aiming to use cardboard.

METHODOLOGY

Materials used to build the hand include cardboard, toothpicks, glue, scissors, rubber bands, nylon cord, an Arduino board, servo motors, wires, and breadboard, all of which were used to construct a low-cost prosthetic hand.

The production of the prosthetic hand began with the development of each digit. Each digit contained double-layered cardboard sides, a toothpick at the upper side of the digit that kept the nylon cord in place, and a toothpick at the bottom, serving as an axis between digits. The nylon cord acted as a tendon, which retracted when it was pulled on. Subsequently, the production of the palmar and dorsal sides of the hand consisted of 2 cardboard frames with holes to tie the strings that held the fingers together. Along with this, 5 nylon cord strands ran through the palm and into the forearm. The forearm consisted of two servos; the fingers were tied onto the servo motors, devices with output shafts were divided to achieve the highest efficiency. These shafts undergo angular changes that are dependent on the coding signal. In contrast, stepper motors were not used; they contain high pole counts resulting in higher torque without the need for an encoder. However, servo motors were used since they run more efficiently and generate torque at higher speeds, thus being more suitable for the project goal (Lackey, 2018). The two servos and the EMG muscle sensor were connected to an Arduino board located inside the forearm. EMG was used as an inexpensive method to control the hand and use existing muscles and muscle memory of the amputee allowing for ease of use, functionality, and aptitude.

The second portion of this research project includes the coding aspect. A threshold was set to ensure the servo motor would only rotate under the condition that the muscle sensor detected a strong signal from the muscle. The servo motors only rotated 100 degrees, which prevented overstretching of the nylon cord, while still allowing enough rotation for the finger to move. This was coded through the Arduino IDE software.

After construction and coding, testing for efficiency and success was required. These tests included the time the fingers take to perform a full range of motion, the range of motion of the fingers, and the strength of finger flexion. The weight and cost of constructing the cardboard prosthetic hand were collected to ensure the hand fit the need of low-cost and low-weight. The prosthetic needed to have a similar function to existing, non-cardboard prosthetics, which was determined through statistical analysis. The speed test was conducted using a timer with the unit of seconds and undergoing three trials per finger. The timer started upon the movement of the finger and ended upon the contact to the palm. This is determined by dividing the range of motion of the MCP joint by the seconds it took to complete a full range of motion. The range of motion tests occurred using a goniometer, and degrees served as the unit of measurement. The ROM was measured from the three joints in each finger (MCP, PIP, DIP). The strength at fingertip test will take place by applying weight to each fingertip, and measure the grams of weight upon the succession of lift from 3 trials per finger.

Following the completion of testing statistical analysis was used to determine if there was a significant difference in the speed, ROM, and strength between the cardboard and non-cardboard hands. The weight and cost was also compared using the averages of both. Testing range of motion is to provide insight into the flexibility of the joint function.

DATA AND RESULTS

To determine the success of the Cardboard prosthetics hand; the weight, cost, range of motion, speed, and strength at fingertip tests were used. Each of the test result's averages were then compared to existing commercial prosthetic devices using statistical analysis to determine if there is a significant difference.

The cardboard prosthetics range of motion success was determined by comparing data referenced from Love et al., consisting of the average active range of motion during finger flexion for the metacarpophalangeal joint (MCP), the proximal interphalangeal joint (PIP), and the distal interphalangeal joint (DIP) for non-cardboard prosthetic hands (2009).

The measurements from the prosthetic hand reference showed an average range of motion of 116° for the MCP joint, 110° for the PIP joint, and 74° for the DIP joint. Whereas, the cardboard hand averaged 111° for the MCP joint, 94° for the PIP joint, and 75° for the DIP joint.

Statistical analysis between the reference and the cardboard hand indicated that the range of motion of the MCP, PIP, and DIP joints in every finger were statistically insignificant compared to the range of motion data collected on human hands. The research hypothesis was tested through t-tests as a form of statistical analysis. The calculated T-values were then compared to the critical value of 3.182. The tests revealed that the range of motion of the MCP, PIP, and DIP joints in every finger were statistically insignificant compared to the range of motion data collected on human hands.

Extended comparisons are made in regards to the cardboard hand against a human hand. Figure 1 demonstrates the insignificant difference in the range of motion between the existing recorded data of the human hand, and the cardboard prosthetic hand, with the joint type being the X-axis, and the range of motion of the Y-axis. The insignificance is shown through the error brackets representing two standard errors of the mean. Since each value for every joint of the prosthetic hand falls under the corresponding human joint's standard error of the mean, the prosthetic hand is statistically similar to the data extracted from the human hands, referenced from Bain et al.

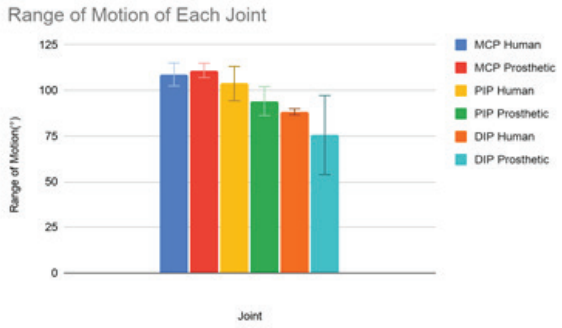


Figure 1

The raw data extracted from the cardboard prosthetic hand's speed test showed the small finger outperformed the other fingers with an average of 2.08 seconds, while the thumb and middle finger had the slowest speeds with an average of 3.05 and 2.95 seconds. This data is conducted to compare the cardboard prosthetic hand's speed to existing prosthetic hands.

Following the cardboard hand speed test, a comparison is made to existing non-cardboard prosthetic hands, from Belter et al., measuring speed performance on prosthetic hands (Belter, Segil, Dollar, & Weir, 2013). Thus, a comparison of the cardboard and commercial prosthetic hands in degrees/second shows the average degrees/second of the commercial prosthetic hand was much larger than the cardboard hand's average, with a value of 78.27 compared to a value of 43.70.

The statistical analysis between averages of degrees/second for the cardboard prosthetic hand and the existing prosthetic hand indicated the rejection of the research hypothesis. This was shown as the t-value of 9.91 was over our critical value of 2.77.

The third test conducted towards the built cardboard prosthetic hand is the strength at fingertip test. From the data, averages from strongest to weakest are index, ring, small, middle, and thumb.

A comparison of the cardboard built hand to existing commercial prosthetic hand shows shows a higher average of force for the commercial prosthetic hand, compared to the cardboard hand. The data for the prosthetic hand from Belter et al. consists of the average fingertip force from the little, ring, index, middle, and thumb fingers (Belter, Segil, Dollar, & Weir,

2013).

Lastly, the statistical analysis of the comparison of the cardboard prosthetic's finger strength and speed to commercially available prosthetic hands concluded that they are significantly different as both t-values exceed the critical values, proving that the cardboard hand failed to replicate the strength or speed of commercial prosthetics.

Lastly, the weight and cost of the cardboard prosthetic hand are recorded to be compared to the average prosthetic hand's weight and cost. The weight of the Cardboard hand is 305 grams, which is lower than the average prosthetic hand weight of 400 grams (Belter & Dollar, 2011). The cost of the Cardboard hand was around $200, while the average cost to make a prosthetic hand ranges from $5,000 to $50,000 (Alliance of Advanced Biomedical Engineering, 2018).

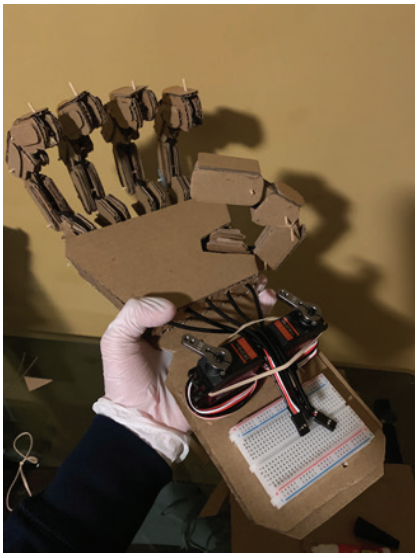


Figure 2

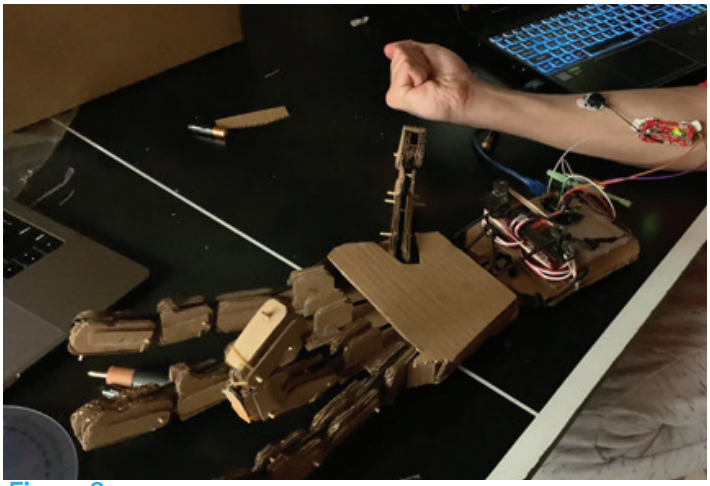


Figure 3

CONCLUSION

The engineering goal stated that the cardboard prosthetic hand would pass all three tests of degrees/second, strength at the fingertip, and range of motion. After testing, the t-test signified the result of a null hypothesis for the ROM test, the alternative hypothesis for the speed test, and the alternative hypothesis for the fingertip force test. In turn, the engineering goal was not met, as the cardboard prosthetic hand had significantly lower fingertip force and degrees/second during finger flexion. Furthermore, despite the weaknesses, cardboard may serve as a viable option in terms of finger movement, to other prosthetic devices. However, it may not be a sufficient substitute for other exterior materials as the cardboard hand lacked force and speed.

However, the adversities of cardboard are apparent in force production, since the prosthetic hand was significantly weaker in fingertip strength and speed of finger flexion compared to commercial prosthetics. This is due to the lack of durability and structural support that cardboard provides compared to the construction present in commercial prosthetics. This may also be due to the servo motors being too weak to rotate quickly enough without breaking surrounding structural support. Another reason for the lack of speed and force may have been the insufficient method of connecting the nylon cords to the server motor, or the fact that there were 2 motors controlling the movement of 5 fingers. This shortcoming is exemplified through the 56% difference in speed between the cardboard hand and the averages of the commercial hands. The speed experimentation also shows a flaw within the fingers, the inconsistency in speed. This is important as this inconsistency may cause the hand to not be able to grasp using all fingers at the same time, this may lead to deeming the prosthetic incapable of emulating humans, or other existing prosthetics ability to grasp.

Additionally, another drawback of the cardboard prosthetic presents itself in force production again. The cardboard hand lacked fingertip strength comparatively to other prosthetic hands. Out of all the tests, this is where the cardboard hand lacked the most. The lack of overall power production may be due to weakened strength and support in many

sub-levels such as the motors, the nylon cord, and the finger construction. Also, there was a clear weakness in the thumb and middle finger of the hand, which is due to the nylon cord sitting loose in the middle finger, resulting in a weaker contraction, and may be solved by increasing the number of motors or space on the servo motor for all of the nylon cords. Additionally, the poor assembly of the nylon cords tying to the servo motor resulted in a loss of power and speed. For the thumb, the weakness is likely due to the lack of structural support rooting from the hand to where the thumb inserts, and the length of the thumb that displaces the weight further away from the structural support of the palm.

The engineering goal is not met, thus there are still many ways to improve the hand as in the use of a prosthetic. Further enhancements such as precision, speed, and durability could be made to allow the prosthetic to function similarly to human hands, which makes the hand more applicable to real-life usage. Each motor could also be signaled from different muscles to more precisely control the functions of the hand. The muscle sensor used during experimentation was subject to interferences especially with the use of surface electrodes, thus intramuscular electrodes would be a sufficient alternative. To ensure a more realistic application of the prosthetic, the materials used need to be more durable while maintaining low cost, so 3D-printed nylon filament would act as an alternative. Through these adjustments, the hand will become more sturdy allowing the hand to be able to carry more weight, however further research into a stronger motor that allows for a full rotation while keeping a compact size is necessary to make the hand stronger overall.

The hand did not succeed or pass all of the engineering tests, the prosthetic lacked some impactful qualities of a human hand, lessening the ability for the hand to act as a viable replacement for commercially available prosthetic limbs. These shortcomings present themselves through lack of full movement of all joints, lack of precise control, lack of force production, and lack of durability. However, the prosthetic hand achieved finger flexion as a response to voltage data collected from the forearm muscles, while emulating similar ranges of motion of a human and non-cardboard prosthetic hand, lower weight, and speed.

Furthermore, the qualities and aspects achieved by the cardboard prosthetic hand, such as use of EMG and cheap materials, can be implemented into future prosthetic devices to aid in lowering the cost and production of prosthetics for amputees.

REFERENCES

Amputation. (n.d.). Retrieved October 24, 2020, from www.hopkinsmedicine.org/health/treatment-tests-and-therapies/amputation

Bain, Gregory & Polites, N & Higgs, B & Heptinstall, R & McGrath, A. (2014). The functional range of motion of the finger joints. The Journal of hand surgery, European volume. 40. 10.1177/1753193414533754.

Belter, J. T., & Dollar, A. M. (2011, June 29). Performance Characteristics of Anthropomorphic Prosthetic Hands. Retrieved from www.eng.yale.edu/grablab/pubs/Belter_ICORR2011.pdf

Belter, J. T., MS, BS, Segil, J. L., Dollar, A. M., Ph.D., SM, BS, & Weir, R. F., Ph.D. (2013, November 5). Mechanical design and performance specifications of anthropomorphic prosthetic hands: A review. Retrieved February 27, 2021, from www.eng.yale.edu/grablab/pubs/Belter_JRRD2013.pdf

Carrozza, M. C., Massa, B., Micera, S., Lazzarini, R., Zecca, M., & Dario, P. (2002, July). The Development of a Novel Prosthetic Hand—Ongoing Research and Preliminary Results. Retrieved October 23, 2020, from www.researchgate.net/publication/3414844_The_development_of_a_novel_prosthetic_hand_-_Ongoing_research_and_preliminary_results

Chan, M., Dr., & Zoellick, R. B., Mr. (2011). World Report On Disability. Retrieved October 23, 2020, from https://www.who.int/disabilities/world_report/2011/report.pdf

Children and its reliability. Retrieved November 20, 2020, from bmcmusculoskeletdisord.biomedcentral.com/articles/10.1186/1471-2474-15-199

Dally, C., Ritter, S., Canon, M., Johnson, D., & Mehta K. (2015, October). Characteristics of a 3D-printed prosthetic hand for use in developing countries. Retrieved October 23, 2020, from www.researchgate.net/publication/285591609_Characteristics_of_a_3D-printed_prosthetic_hand_for_use_in_developing_countries

D'Ausilio, A. Arduino: A low-cost multipurpose lab equipment. Behav Res 44, 305–313 (2012). doi.org/10.3758/s13428-011-0163-z

Electromyography (EMG). (n.d.). Retrieved February 10, 2021, from www.hopkinsmedicine.org/health/treatment-tests-and-therapies/electromyography-emg#:~:text=Electromyography%20(EMG)%20measures%20muscle%20response,nerve's%20stimulation%20of%20the%20muscle.&text=EMG%20measures%20the%20electrical%20activity,produce%20electrical%20signals%20during%20rest.

Lackey, B. (2018, June 24). What's the Difference Between Servo and Stepper Motors? Retrieved October 23, 2020, from www.machinedesign.com/mechanical-motion-systems/article/21836868/whats-the-difference-between-servo-and-stepper-motors.

Lam, S. (n.d.). Myoelectric Prosthesis. Retrieved October 23, 2020, from bme240.eng.uci.edu/students/10s/slam5/considerations.html

Love, L., Lind, R., & Jansen, J.F. (2009). Mesofluidic actuation for articulated finger and hand prosthetics. 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2586-2591.

Merrett, G. (n.d.). Southampton Hand Assessment Procedure. Retrieved October 23, 2020, from www.shap.ecs.soton.ac.uk/about.php

Nguyen, N. T. N. (2018, May 11). Developing A Low-cost Myoelectric Prosthetic Hand. Thesis. https://www.theseus.fi/bitstream/handle/10024/147196/thesis_susan.pdf?sequence=1&isAllowed=y.

Raez, M. B., Hussain, M. S., & Mohd-Yasin, F. (2006, March 23). Techniques of emg signal analysis: Detection, processing, classification and applications. Retrieved February 10, 2021, from www.ncbi.nlm.nih.gov/pmc/articles/PMC1455479/

Raichle, K., Hanley, M., Molton, I., Kadel, N., Campbell, K., Phelps, E., . . . Smith, D. (2008). Prosthesis use in persons with lower- and upper-limb amputation. Retrieved October 24, 2020, from www.ncbi.nlm.nih.gov/pmc/articles/PMC2743731/.

Statistics on hand and arm loss. (2014, February 4). Retrieved October 23, 2020, from www.ishn.com/articles/97844-statistics-on-hand-and-arm-loss

Talbot, K. (2014, May). Using Arduino to Design a Myoelectric Prosthetic. Retrieved October 23, 2020, from digitalcommons.csbsju.edu/cgi/viewcontent.cgi?article=1052&context=honors_theses

3-D printing to lower prosthetic costs. (n.d.). Retrieved March 08, 2021, from aabme.asme.org/posts/3-d-printing-to-lower-prosthetic-costs

Vanderesa, J. (n.d.). Arm & Hand Prosthetics. Retrieved October 23, 2020, from mcopro.com/blog/resources/arm-hand-prosthetics/

Vasluian, E., Bongers, R. M., Reinders-Messelink, H. A., Dijkstra, P. U., & Van der Sluis, C. K. (2014, June 10). Preliminary study of the Southampton Hand Assessment Procedure for

The Early Detection of Type 2 Diabetes Using Machine Learning Algorithms

Shridha Rajeswar
duPont Manual High School
11th, Louisville, Kentucky

ABSTRACT

When it comes to Type 2 Diabetes, it can be difficult to derive accurate diagnosis results with current invasive screening tests that provide false-positive and false-negative results. This issue can be mitigated with the help of machine learning (ML) algorithms that can accurately predict a patient's diabetic condition. It has revolutionized many industries, like information technology and finance, and healthcare. Among the different ML types, three classification algorithms were used to classify input values into output values that signify one's diagnosis result for Type 2 Diabetes.

The engineering goal of this project was to develop several ML algorithms (K-Nearest Neighbors, Logistic Regression, and Decision Tree) that will accurately output diabetic results using a binary number system (0 for not diabetic, 1 for diabetic) based on seven types of vital body measurements that indicate diabetes (body mass index, skin thickness, etc.). The algorithms were developed using Python and the sample data were derived from a Kaggle database.

There were different independent variables for each type of ML algorithm due to the differences in frameworks, but they were chosen to determine their effect on the accuracy of the predicted output results. It was generally observed that a stable number of nearest neighbors (15) for KNN, learning rate (0.3) for LR, and maximum depth (4) for DT produced the most accurate algorithms within each algorithm type. However, there were a few variables (number of iterations and minimum number of samples for node splitting) that did not seem to affect the accuracy rates, and statistical analysis tests led to the acceptance of the null hypotheses. Using the derived results, more efficient algorithms can be implemented in a potential detection hardware prototype that would be used to test the practicality of a possible ML diagnosis device.

Keywords: type 2 diabetes, python, machine, learning, k-nearest neighbors, logistic regression, decision tree, training, predicting

INTRODUCTION

Type 2 Diabetes is a condition that around 380 million people around the world live with daily (Santos-Longhurst). According to MedlinePlus, "Insulin is a hormone that helps glucose get into your cells to give them energy" (2 Sept. 2020). People with type 2 diabetes tend to be insulin-resistant; when insulin — the hormone that breaks down glucose — is resisted, the body's glucose builds up in the bloodstream, leading to insulin overproduction from the pancreas. Eventually, the insulin-producing cells soon weaken and thus do not produce enough insulin (26 Aug. 2020). Such a condition would need to be detected early to prevent complications as diabetes can lead to many problems such as heart disease, neuropathy, vision/hearing impairments, and slow healing.

Several early detection screening tests are used on patients, however, they can often get inaccurate results. This is because "there are no randomized clinical trials documenting the effectiveness of screening programs in decreasing mortality and morbidity from diabetes, and some controversy exists regarding the cost-effectiveness of screening and whether screening as currently carried out is a systematic and ongoing process" ("Screening for Type 2 Diabetes"). There is also the possibility of having false-positive and false-negative results, which would underestimate or overestimate the likelihood of someone having type 2 diabetes. Research indicates little to no information on how or why certain people's bodies become insulin-resistant, which makes it even more difficult to get accurate screening results as specialists wouldn't know what to expect as the result.

However, technological advancements such as artificial intelligence could help doctors

and researchers in detecting diabetes more accurately in patients and possibly even in finding a cure. Machine Learning (ML) is defined as the "field of study that gives computers the capability to learn without being explicitly programmed" ("Machine Learning"). It allows machines that implement this concept to carry out tasks and learn patterns at the same time, and by doing so, machines can expand their knowledge and carry out commands without the instruction of external users. ML algorithms are programs that are used by devices and machines to accomplish tasks through artificial intelligence (14 May 2020). Machines that use such algorithms would learn like a human and gain experience over time, eventually getting faster and more accurate at tasks and commands. With the increasing accuracy, ML screenings could provide more accurate results, thereby increasing the chances of early diagnosis and reducing the complications of diabetes.

In the algorithms that will be developed, there are two phases: training and predicting. In the training phase, several diabetes indicators such as age, glucose levels, and so on will be stored in a training dataset, while the rest of the parameters are used as input values. These input values are passed through the network to obtain preliminary predictions. The predicted values are then compared with the expected values from the training dataset to calculate the error (the difference between the expected output and the actual output) and accordingly, the predictions are altered to produce new output values. After some training, the algorithm would enter the predicting phase, where it produces the final output values for the other input values that were fed to it.

K-Nearest Neighbors (KNN) is an ML algorithm that is commonly used to solve classification and regression tasks. KNN behaves similarly to humans; people are affected by the people in their life and their surroundings, and KNN is also affected by parameters that have neighboring or close values (Brownlee, 2020). By using these parameters, this algorithm classifies the data into different clusters based on their proximity to one another. The independent variables that will be altered to produce different KNN algorithms are number of folds, which determines the number of data present in each cluster, and the number of nearest neighbors, which determines how many nearest neighbors can be used in assessing a particular data point's cluster (and which will be used to create different experimental groups).

Logistic Regression (LR) is another ML algorithm that is also useful in helping with regression problems. This type of algorithm uses a function called sigmoid to predict the output values given a set of input data and to reduce the computational errors that may occur along the way (Pant, 2019). The independent variables that will be altered to produce different LR algorithms are the number of iterations, which is the number of times that the algorithm got trained with the given training dataset, and learning rate, which will be used to create different experimental groups ("Sklearn.linear_model.LogisticRegression").

Decision Tree is the third machine learning algorithm that is used in recognizing patterns, which would therefore allow it to make predictions about behaviors in data. This is the type of algorithms used by autonomous cars to self-drive using various algorithms to predict the behavior on roads and terrains as well as at signals (Yse, 2019). Such an algorithm is ideal for solving classification and regression problems and is structured like the branches of a tree. The independent variables that will be altered to make different Decision Tree algorithms will be depth, which affects how deep or extensive the algorithm is, and the minimum number of samples for node splitting, which affects how nodes will split based on the number of samples in each group (and which will be used to create different experimental groups).

For this project, the engineering goal is to develop ML algorithms that would use the input of eight diabetic parameters to determine if a person has diabetes or not using binary classification as well as to determine which algorithm would most accurately predict the early detection result. The algorithms will be coded using Python. The independent variables will vary across each algorithm, and the dependent variables will be the accuracy rates. To get unbiased results, random diabetic parameter sample data derived from a Kaggle database will be used to train the algorithm. The accuracy rates of the machine learning programs will be determined by calculating the percentage of

correct output values over the total percentage of output values, and the significance between the accuracy rates will be compared using the t-test.

METHODOLOGY

The three machine learning algorithms were coded and their respective independent variables were varied. Each variation was then allowed to learn, and the final accuracy rate was recorded. All of the information and processes were learned from a programmer's website called Machine Learning Mastery.

DATA AND RESULTS

The inputs for the algorithm were the diabetic parameters for detecting Type 2 Diabetes (plasma glucose concentration, blood pressure, skin thickness, insulin levels, body mass index, diabetes pedigree function, and age), and the outputs were binary numbers that indicate whether the unidentified patients had diabetes or not (0 = non-diabetic, 1 = diabetic). There were multiple independent variables across each type of machine learning algorithm: number of folds and number of nearest neighbors for KNN, learning rate and number of epochs for LR, and maximum depth and minimum number of samples for node splitting for DT. A total of 6 KNN algorithms, 9 LR algorithms, and 9 DT algorithms were tested out, and a total of 2, 3, and 3 experimental groups were tested within each type of algorithm, respectively.

Figure 7 depicts the accuracy rates for the KNN algorithms at different numbers of folds and numbers of nearest neighbors. Figure 8 shows the accuracy rates for the Logistic Regression algorithms with differing learning rates and number of iterations (or epochs). Figure 9 illustrates the trends for the accuracy rates of the Decision Tree algorithms at different maximum depths and minimum numbers of samples for node splitting. Figure 10 sums up most of the observed patterns by depicting the average accuracy rates of the most accurate algorithms in each algorithm type group. In other words, the accuracy rates of the algorithms in the most effective experimental group of each algorithm type were averaged and compared to one another to determine the most accurate program across all three algorithms. This was done to understand which independent variable values were most ideal for the algorithms that had higher accuracy rates than the rest of the programs. Figure 11 depicts the t-test outcomes for four t-tests, with one of the t-tests coming out as inconclusive due to the absence of variation.

In general, it could be seen that for the Logistic Regression and Decision Tree algorithms, the algorithms that had higher accuracy rates (and thus more efficient predicting capabilities) when the independent variables were right in the middle of the chosen values; that is, the accuracy rates demonstrated peaks in the middle and decreased on either side of the middle value. For instance, in Figure 8, the LR programs with a learning rate of 0.3 had a higher average accuracy rate of around 78.04% than the programs with learning rates of 0.1 and 0.5. However, this trend does not apply to the KNN algorithms; it can be seen that the lower value of 5 folds yielded more effective programs than the higher value of 10 folds, with accuracy rates at least 0.30% higher.

When comparing the programs within their algorithm type, several deviated trends were not predicted. In the LR algorithms, for example, all the programs with epochs of 1000, 5000, and 20000 had the same accuracy rates of 77.65%, 78.04%, and 77.65%, respectively. It can be thus inferred that the number of epochs or iterations doesn't have a significant impact on the efficiency of the LR programs. Likewise, the minimum number of samples for node splitting in the DT algorithms didn't seem to have much of an effect on the accuracy rates; the programs with minimum samples of 5, 7, and 9 had the same accuracy rates of 73.33%, 74.38%, and mostly 74.12%, respectively. However, this was not the case with the KNN algorithms. Both the number of folds and the number of nearest neighbors had impacted the effectiveness of the programs. Due to these observations, future adjustments could be made so that the number of folds, which were kept constant as 5 folds in the LR and DT algorithms, could be changed to see if the number of folds plays a role as an independent variable for changing accuracy rates.

Though the goal of the study was to compare the accuracy rates of different types of algorithms and a hypothesis was not postulated, multiple paired t-tests were still conducted to observe any significant changes between the accuracy rates for each algorithm type. To make things simpler, the accuracy rates of

the most effective experimental group in each algorithm type were averaged to have a single value representing each type of algorithm. After conducting the t-test, it was found that most of the differences in the accuracy rates were not significant and the null hypotheses were accepted for all of them. In Figure 11, for the t-test comparing the most accurate LR algorithm with a learning rate of 3 (mean value = 78.04%, standard deviation = 0%) to the most accurate KNN algorithm with 5 folds (mean value = 74.51%, standard deviation = 1.38%), however, it can be seen that there was a more significant change, with an observed t-value of 4.432 compared to the other t-values of 1.243 and 0.1632 from the other t-tests. For the t-test comparing the most accurate LR algorithm with a learning rate of 3 to the most accurate DT algorithm with a maximum depth of 4, the test was inconclusive because the standard deviation values were zero, indicating that the accuracy rates within the experimental groups were very consistent with no variation.

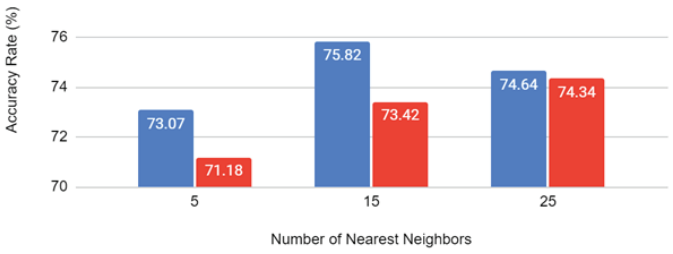


Figure 7: Display of the accuracy rates of algorithms in two different experimental groups based on the number of folds (KNN)

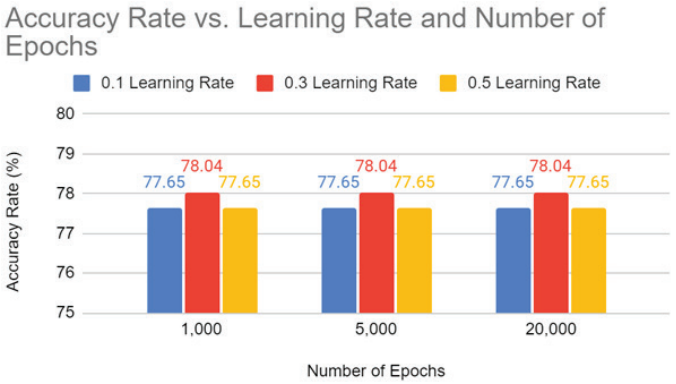


Figure 8: Display of the accuracy rates of algorithms in three different experimental groups based on the learning rate (LR)

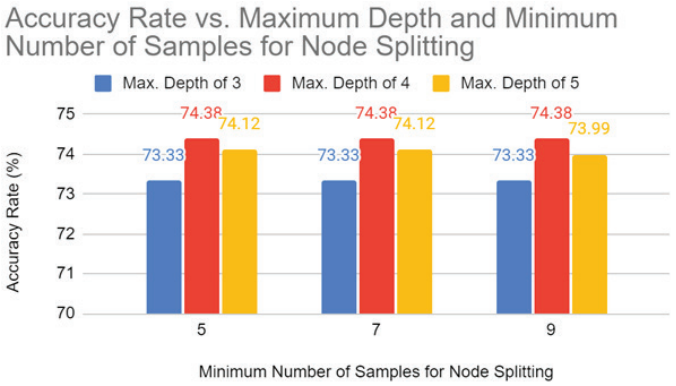


Figure 9: Display of the accuracy rates of algorithms in three different experimental groups based on the maximum depth (DT)

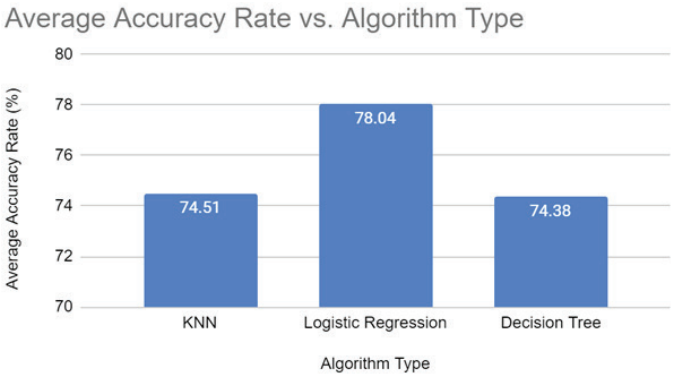


Figure 10: Display of the average accuracy rate of the algorithms in the most effective experimental groups for each type of algorithm (5 folds for KNN, 0.3 learning rate for LR, and maximum depth of 4 for DT)

Comparison (1 vs. 2)	Mean 1	Mean 2	Variance 1	Variance 2	T-value (observed)	T-value (0.005 level)	Hypothesis
5 folds vs 10 folds (KNN)	74.51	72.98	1.903	2.642	1.243	9.925	Null accepted
0.3 learning rate vs. 5 folds	78.04	74.51	0	1.903	4.432	9.925	Null accepted
5 folds vs maximum depth of 4	74.51	74.38	1.903	0	0.1632	9.925	Null accepted
0.3 learning rate vs maximum depth of 4	78.04	74.38	0	0			Inconclusive (no variation)

Figure 11: Table showing the paired t-test results for the programs with the highest average accuracy rates in each algorithm type

CONCLUSION

In this study, three types of machine learning algorithms were assessed on their accuracy of predicting the diabetic conditions in patients. This was done by providing anonymous data values indicating diabetic parameters (blood pressure, insulin levels, and so on) to the algorithms and training them in order to have them predict the diabetic conditions of new sample data later on in real-time. For each type of program, different experimental groups were created based on the manipulation of different independent variables, and each program was run 10 times. Later, the accuracy rates were determined by calculating the number of results that were predicted correctly over the total number of results predicted, and these rates were compared across experimental groups and the three types of algorithms.

Most of the accuracy rates within each experimental group were found to be the highest when the independent variables were in the middle range. In LR algorithms, a lower learning rate results in an algorithm that trains slowly but also predicts more efficiently over time, and a higher learning rate results in an algorithm that trains quickly and predicts with more errors. These two circumstances would not be ideal for predicting Type 2 Diabetes, as it could lead to prediction errors and resulting in false negative and false positive diagnoses that are expensive in terms of time. However, with a middle range learning rate, the algorithms can predict accurate and time-efficient results that would help in diagnosing patients. This result was expected from the previous year's investigation, where it was found that the ideal learning rate for artificial neural networks predicting water quality balanced out time costs and training accuracy. In terms of DT, a larger maximum depth can often lead to the overfitting of a program's predictions, and a smaller maximum depth can lead to the underfitting of an algorithm's predictions. Again, these two situations are not ideal for predicting Type 2 Diabetes as it can lead to false positive and false negative diagnoses, which could ultimately lead to a waste of money and resources and deteriorating health, respectively. Therefore, DT algorithms with middle range maximum depths would be ideal in predicting the diabetic state of a patient as they would predict results that don't overfit or underfit the given input data. This contradicts an assumption in "PRODUCTS AND HELP BITS IN DECISION TREES," where it mentions that "The cost of a computation (decision tree) is simply the number of input variables that are read (the depth of the decision tree)" (Nisan et. al, 1999). Although seven input variables were used to determine the diabetic conditions of sample data, the programs with a maximum depth of 4 had a higher accuracy rate, thus refuting Nisan et. al's statement.

However, this trend did not apply to KNN algorithms possibly because there were only two KNN experimental groups compared to the three experimental groups in the other types. However, this also refutes the trend that some other studies found. For instance, in the study "Improved kNN Algorithm by Optimizing Cross-validation," the authors determine that the optimum number of folds for a KNN algorithm is 10 because it allows for the cross-validation of training and testing predictions in a way that produces the most accurate results (Dadhania and Dhobi). Due to the lack of time, only two values of folds were used to test the accuracy of each experimental group, but if another additional experimental group of 15 folds was tested, it is plausible that the experimental group with 10 folds would have a higher average accuracy rate than the rest of the algorithms.

Overall, when comparing across all types of algorithms based on the average accuracy rate of the most efficient experimental groups, it was concluded that the Logistic Regression algorithms had the highest average accuracy rate of about 78.04%, but it was also found that there was a significant difference in average accuracy rates between the LR algorithms with a learning rate of 0.3. For the t-test comparing the most accurate LR algorithm with a learning rate of 3 to the most accurate DT algorithm with a maximum depth of 4, the test was inconclusive, and this stems from the trend in Figures 8 and 9, where the number of epochs and the minimum number of samples for node splitting did not have any effect on the accuracy rates of the experimental groups in the LR and DT algorithms.

This study is a developing base for future research that could achieve more advanced and sophisticated goals. One improvement would be

to test more values of independent variables to ensure that the trends of the ideal values being in the middle range still apply. This would provide the full scope on the effect of the manipulation of each independent variable on the accuracy rates. Another change would be to actually implement these machine learning algorithms with the help of hardware and gadgets that would detect health conditions in real-time. By possibly using raspberry pi as a base to implement the machine learning software, future studies could help medical professionals and bioengineers in designing a detector machine that uses machine learning to easily and accurately detect health conditions to provide early diagnosis and treatments, which could save patients' costs, time, and most importantly, health.

REFERENCES

Brownlee, J. (2020, August 14). K-Nearest Neighbors for Machine Learning. Retrieved from https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/

Dadhania, S. S., & Dhobi, J. S. (2012). Improved kNN Algorithm by Optimizing Cross-validation. International Journal of Engineering Research and Technology, 1(3), 1-6. Retrieved from https://www.ijert.org/research/improved-knn-algorithm-by-optimizing-cross-validation-IJERTV1IS3186.pdf.

Gujral, S. (2017). Early Diabetes Detection Using Machine Learning: A Review. –International Journal for Innovative Research in Science & Technology, 3(10), 57-62. Retrieved from http://www.ijirst.org/articles/IJIRSTV3I10027.pdf

Machine Learning. (n.d.). Retrieved from http://www.contrib.andrew.cmu.edu/~mndarwis/ML.html.

Machine Learning Algorithms: Top 5 Machine Learning Algorithms. (2020, May 14). Retrieved from http://www.edureka.co/blog/machine-learning-algorithms/

Mitesh, W., Kumar, V., Tarale, S., Galgat, P., & Chaudhari, D. J. (2019). International Research Journal of Engineering and Technology. Diabetes Diagnosis Using Machine Learning Algorithms., 6(3), 1470-1476. Retrieved from www.irjet.net/archives/V6/i3/IRJET-V6I3277.pdf.

Nisan, N., Rudich, S., & Saks, M. (1998). Products and Help Bits in Decision Trees. SIAM Journal on Computing, 28(3), 1035. https://doi.org/10.1137/S0097539795282444

Pant, A. (2019, January 22). Introduction to Logistic Regression. Retrieved from https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. Baltic J. Modern Computing, 5(2), 221-232. Retrieved from https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/5_2_05_Pranckevicius.pdf.

Santos-Longhurst, A. (n.d.). Type 2 Diabetes Statistics and Facts. Retrieved from www.healthline.com/health/type-2-diabetes/statistics#5

Screening for Type 2 Diabetes. (2004, January 01). Retrieved from https://care.diabetesjournals.org/content/27/suppl_1/s11

Sklearn.linear_model.LogisticRegression¶. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Type 2 diabetes. (2020, August 26). Retrieved from http://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-caus es/syc-20351193.

Type 2 Diabetes| Adult-Onset Diabetes. (2020, September 02). Retrieved from https://medlineplus.gov/diabetestype2.html

Yse, D. L. (2019, April 30). The Complete Guide to Decision Trees. Retrieved from https://towardsdatascience.com/the-complete-guide-to-decision-trees-28a4e3c7be14

A Reconceptualization of the Asymmetric Multiple Traveling Salesman Problem (amTSP): Engineering an Approximate Algorithm for Cost

Shubh Gupta
duPont Manual High School
11th, Louisville, Kentucky

ABSTRACT

The asymmetric multiple Traveling Salesman Problem (amTSP) is an NP-hard optimization problem in which hamiltonian cycles of least cost are to be found. Given a set of nodes, the objective is to find an optimized tour in which each node must be covered once by one salesman and all salesmen return to a starting node. Countless heuristic and exact methods have been researched in the past, however, they cannot output a projected cost when placements of nodes are not known. This study sought to develop such an amTSP estimation algorithm that takes the number of nodes, area of the region containing the nodes, and the number of salesmen, to output a least cost accurate to 1%. An amTSP solver was first created in MATLAB utilizing a genetic algorithm. Trials were conducted for each set of nodal and salesmen values and area was based on the smallest polygon enclosing the nodes. The study concluded that radius and cost showed a directly proportional relationship, nodes and cost showed a strong exponential relationship, and salesman and cost per salesmen had a weak inverse relationship. Several supervised learning methods were used in developing an estimation algorithm. The Gaussian Process model was found to be most effective, with a percent error of 7.2%; however, it failed to reach the engineering goal. This study also reveals the potential of a deep learning model to exceed the Gaussian Process and sparks novel research into a new field in optimization: amTSP planning and estimation.

INTRODUCTION

As society continues to accelerate through the 21st century, efficiency has become of utmost importance. This urge to make the most of a small amount of time has thrust mathematical optimization problems into the spotlight, especially the Traveling Salesman Problem (TSP). The TSP is concerned with finding "the shortest path [or least cost] that a salesman should take to traverse through a list of cities [or nodes]" while returning back to the first node (Ataee, 2020). TSP is said to be NP-hard, which means that no algorithm has been developed to solve all instances of the problem in a reasonable time. In fact, a reasonable-time solution for the problem earns a $1,000,000 prize from the Clay Mathematics Institute (Cook, 2015).

This research will be focused on one specific variation of the traditional TSP: the asymmetric multiple Traveling Salesman Problem (amTSP). In this version, the distance from node a to node b is not necessarily equal to the distance from b to a (asymmetric). This simulates real-world complications such as traffic and roadblocks. Additionally, more than one salesman is used to visit each node (multiple). Each salesman is defined to start and finish at a central hub node and each node must only be visited by one salesman. The objective is to minimize the total distance traveled by all the salesmen (Weisstein, n.d.). The amTSP more accurately simulates real-world delivery problems that a company such as UPS would encounter daily. Many exact and heuristic methods have been developed to solve the amTSP based on distances between cities which are compiled in a distance matrix (Cook, 2015).

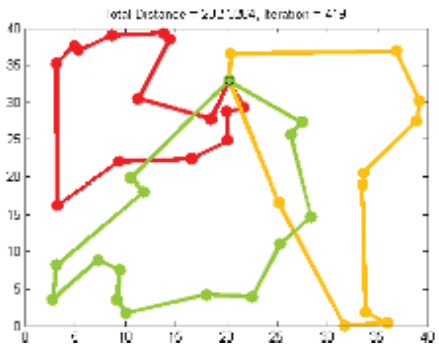


Figure 1: Visual Depiction of the Problem Statement of the multiple TSP

Previous research on the amTSP has been focused on formulating an algorithm or procedure for solving (taking in the distance matrix or location of nodes to output a path of least cost) instances of the problem. These can be broken down into two approaches: heuristics and exact methods. Notable heuristic methods such as tabu search, ant colony optimization, and genetic algorithms aim to find a "close-enough" solution in a short amount of time. Each of the aforementioned methods has well-researched applications to many optimization problems and has been modified to fit the amTSP. Exact methods, such as branch-and-bound, systematically consider all possible routes to find the optimal one. Often, these algorithms run in unreasonable time, however, in recent years, for low nodal values they often run relatively quickly.

Although these algorithms can compute a path of least cost given the distance matrix, they cannot output a projected cost when the locations of nodes are unknown. For instance, consider a local start-up that based on predictive analytics knows how many daily packages they have to deliver in a certain time frame across a certain region. They know the area of the region, the number of orders, and the number of "salesmen" they will have available but do not know the exact locations of the orders. If they wish for an estimated cost based on these factors, perhaps to determine the time it will take or if the route is feasible, an algorithm must be developed to output cost given those aforementioned factors. Currently, no amTSP algorithm with an indefinite distance matrix exists (Bektas, 2006). The purpose of this project is to develop such an estimation algorithm, the first of its kind, that will allow for an accurate preliminary cost that can be used in planning for any application of the TSP. These applications range from vehicle routing to scheduling meetings to positioning cylinders in a printing press (Bektas, 2006). Also, such an algorithm can be used as a standard for an amTSP solver to help determine whether or not a route can be optimized further.

Therefore, the question asked in this inquiry is, how does the placement of nodes in the amTSP impact the total minimum cost of traversing all of them with multiple paths? The engineering goal is to develop an amTSP estimation algorithm that takes in the number of nodes, the number of salesmen, and the area of the region covered by the nodes (independent variables) to output a minimum cost (dependent variable) that is accurate to 1%. This algorithm will be tested by generating 100 random sets of nodes and number of salesmen and comparing the estimated cost to the actual cost. The actual cost will be determined by a metaheuristic method such as ant colony optimization or a genetic algorithm.

The need for this research is even more apparent in the face of the COVID-19 pandemic, where this algorithm could help in planning the logistics of delivering vaccines. Additionally, this project has the potential to spark further inquiry into the unexplored field of TSP estimation algorithms to maximize efficiency and overall, change the world for the best.

METHODOLOGY

This project was conducted on a home iMac, as its 3.2 GHz, Intel Core i5 processor and sufficed in running MATLAB R2020b, the target programming environment. This algorithm was tested by generating 100 random sets of number of salesmen, number of nodes, and region size, and then comparing the algorithm's estimated cost to the actual cost. The actual cost was determined by the genetic algorithm. However, the most intensive step was initially developing the estimation algorithm.

Firstly, the mTSP genetic algorithm (GA) developed in the prior year's project (originally influenced by Joseph Kirk's code) was downloaded onto the registry. Although GA is a heuristic method, it has been proven to be effective in formulating an accurate estimation algorithm in prior research and is relatively easy to manipulate. This code was modified to suit the amTSP definition by relaxing constraints and modifying variables. Also, the distance matrix which stores distances between inputted points was to be scaled so that the distance from point a to b varies slightly from the distance b to a.

Additionally, a separate amTSP point generator program was created in the main.m function to generate the points that were input into the GA. A random xy-coordinate pair (ranging from 5km to 50km for x and y) can be assigned to each node and then compiled into a distance matrix (dmat) to input into the GA. Finally, a method that calculates the area of the smallest polygon enclosing all the points was created by utilizing the built-in "boundary (x,y)" method. Although not exact, this was deemed

the region size for the purposes of this project. Since the area represents the true area of the region, it was deemed to have units of square kilometers. Consequently, since cost is derived from the distance matrix as well, it had an implied unit of km, however, the asymmetric nature of the problem definition can lead to other interpretations of cost (such as work or time), thus units will be left off.

Both these programs were used collectively to generate thousands of sets of points of varying values of number of salesman (ranging from 1-5), number of nodes (ranging from 20-100), and region area paired with the reported least cost by the GA. All of this data was inputted into a Microsoft Excel spreadsheet which was transferred to the MATLAB Regression App. Here, all of the available machine learning regression algorithms were applied to find the best-fit model. A deep learning model was also employed and compared against the machine learning regressions. This deep learning model consisted of four layers: a three-parameter input layer, a fully connected layer, a rectified linear unit layer, and a regression output layer. Finally, the procedure to test the accuracy of the deep learning and regression algorithms mentioned earlier in the methodology was applied and a determination was made on whether or not the engineering goal was met.

DATA AND RESULTS

To analyze the relationship between each of the independent variables (salesmen, nodes, and area) and the dependent variable (cost), the following figures were constructed from an average of cost within the given independent variables. While these trends were not the focus of this study, their determination may help in guiding future research in further exploring their specifics.

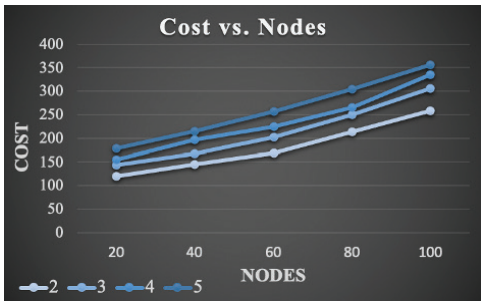


Figure 2: Cost vs. Nodes for all Salesmen Values

The figure above plots average cost against the number of nodes. Area was omitted in this figure since it was a randomized control. A clear exponential relationship can be seen between cost and nodes for all salesmen values. Specifically, exponential regression on the four curves yields r-squared values of higher than 0.98. This indicates with high certainty that the number of nodes follows an exponential pattern indefinitely.

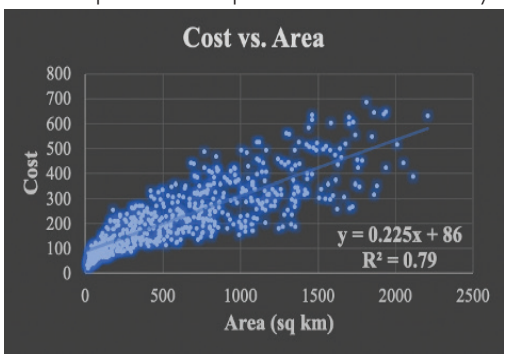


Figure 3: Cost vs. Area

Figure 3 displays the average cost plotted against the area of the region encompassing the nodes for all salesmen and nodal values. A clear linear relationship emerges with the equation displayed on the graph achieving an r-squared value of 0.79, indicating a strong direct relationship. The 0.225 slope signifies that each increase of one sq km will result in a 0.225 increase in cost.

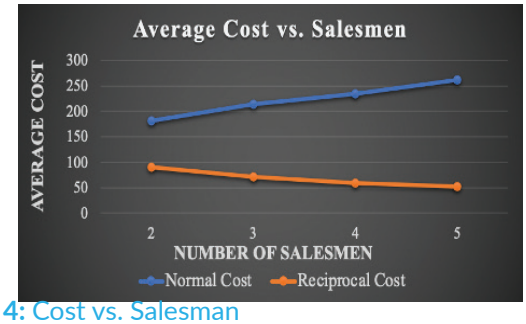


Figure 4: Cost vs. Salesman

Finally, figure 4 shows cost plotted against salesmen (in blue) and the reciprocal cost—or average cost per salesman (in orange). While the data wasn't sufficient in determining the relationship, the average cost clearly increased as more salesmen were added. On the other hand, the reciprocal cost decreased steadily.

After these three relationships were evaluated, the full data could now be considered to design an approximate algorithm for cost that would take in a set of independent variables and output cost. The figure below displays the six tests conducted through MATLAB's supervised learning regression app. 100 of the 600 total data points were withheld from the algorithm to validate and calculate a root mean square error (RMSE).

1 ☆ Gaussian Process Regression Last change: Optimizable GPR	RMSE: **24.068** 3/3 features
2 ☆ Linear Regression Last change: Linear	RMSE: 46.438 3/3 features
3 ☆ Ensemble Last change: Optimizable Ensemble	RMSE: 29.091 3/3 features
4 ☆ SVM Last change: Optimizable SVM	RMSE: 130.42 3/3 features
5 ☆ Tree Last change: Optimizable Tree	RMSE: 36.604 3/3 features
6 ☆ Stepwise Linear Regression Last change: Stepwise Linear	RMSE: 39.542 3/3 features

Table 1: List of MATLAB **Regression App Tests Conducted**

With an RMSE of just 24, the Gaussian Process Regression (GPR) was found to be the most effective model, translating to a 7.2% error and r-squared value of 0.97. Lastly, a deep learning model was also trained, however, it failed to surpass the GPR with an RMSE of 29.0.

CONCLUSION

The purpose of this study was to develop an estimation algorithm, the first of its kind, that will allow for a predicted cost, to be used in planning, for any application of the amTSP. The engineering goal stated that the algorithm was to take in the number of nodes, the area of the region (that contains the nodes), and the number of salesmen to output an estimated cost of the tour accurate to 1%.

Firstly, the study found a direct exponential relationship between nodes and cost, supporting the determination made with a smaller sample size in the prior year's study on the mTSP. However, in the grander scheme of literature on the TSP, this relationship remains inconclusive (Uğur, Korukoğlu, Çalıskan, Cinsdikici, & Alp, 2009). While pursuing this relationship may initially seem like a valuable future area of research, it will likely never be proven rigorously since adding one additional node may result in a completely different optimal tour.

Additionally, this year's study provided the first analysis of the relationship between cost and the area of the region containing the nodes. The two variables were found to have a direct linear relationship with a linear regression equation of y=0.225x+86. The 0.225 slope signifies that each increase of one sq km will result in a 0.225 increase in cost. However, this is a prime target for future research through a pure mathematical approach.

The final relationship investigated, salesmen and cost, was inconclusive; however, a clear increase in total cost was seen as the number of salesmen increased and a decrease in the average cost. This indicates that while utilizing multiple salesmen or vehicles may lead to higher overall expense, it will lead to a decrease in overall time taken. This fact will prove incredibly useful for companies that encounter vehicle routing problems (VRPs) daily. The GPR algorithm developed in this study will reflect this and allow users to maximize productivity.

GPR, the same supervised learning regression approach that was the most effective in last year's examination of the mTSP, remained the best estimation algorithm with an RMSE of 24 and percent error of 7.2%. In all, while the approximate algorithm for cost engineered in this study did not meet the 1% benchmark, it still holds incredible promise for any individuals and organizations seeking to streamline their route planning and maximize efficiency by having an accurate estimation of cost. Additionally, this work paves the way for future investigation into the amTSP, especially with regards to uncovering the relationship between nodes and cost and nodes and salesmen as well as the efficacy of a deep learning model. Since this work has been the first all-inclusive attempt at such an algorithm, it has the potential to spark a new field of research: amTSP estimation and planning.

REFERENCES

Ataee, P. (2020, June 14). How to Solve Traveling Salesman Problem-A Comparative Analysis. Retrieved October 27, 2020, from https://towardsdatascience.com/how-to-solve-the-traveling-salesman-problem-a-comparative-analysis-39056a916c9f

Bektas, T. (2006). The multiple traveling salesman problem: an overview of formulations and solution procedures. Omega, 34(3), 209–219. doi: 10.1016/j.omega.2004.10.004

Cook, W. (2015, August). The Problem. Retrieved from http://www.math.uwaterloo.ca/tsp/problem/index.html.

Joshi, V. (2017, November 13). The Trials And Tribulations Of The Traveling Salesman. Retrieved October 27, 2020, from https://medium.com/basecs/the-trials-and-tribulations-of-the-traveling-salesman-56048d6709d

Ouyang, X., Zhou, Y., Luo, Q., & Chen, H. (2013). A Novel Discrete Cuckoo Search Algorithm for Spherical Traveling Salesman Problem. Applied Mathematics & Information Sciences, 7(2), 777-784. doi:10.12785/amis/070248

Rostami, A., Mohanna, F., Keshavarz, H., & Hosseinabadi, A. (2015). Solving Multiple Traveling Salesman Problem using the Gravitational Emulation Local Search Algorithm. Applied Mathematics & Information Sciences, 9(2), 1–11. doi: 10.12785/amis/090218

Stolaroff, J. K., Samaras, C., O'Neill, E. R., Lubers, A., Mitchell, A. S., &

Ceperley, D. (2018). Energy use and life cycle greenhouse gas emissions of drones for commercial package delivery. Nature Communications,9(1). doi:10.1038/s41467-017-02411-5

Uğur, A., Korukoğlu, S., Çalıskan, A., Cinsdikici, M., & Alp, A. (2009). Genetic Algorithm Based Solution for TSP on a Sphere. Mathematical and Computational Applications,14(3), 219-228. doi:10.3390/mca1403021

Weisstein, E. (n.d.). Traveling Salesman Problem. Retrieved from http://mathworld.wolfram.com/TravelingSalesmanProblem.html.

The Effect of Ambiguous Illusions on VMW Performance in ADHD Populations

Sonia Nagpal
duPont Manual High School
11th, Louisville, Kentucky

ABSTRACT

This research aimed to answer the question of "How do ambiguous optical illusions affect the visual working memory performance of populations with attention-deficit hyper disorder (ages 18-25)?" The purpose of this research was to study if viewing ambiguous illusions could be an effective, practical tool in improving visual working memory (VWM) performance in young adult ADHD populations, who have VWM impairments. It was hypothesized that the viewing of ambiguous illusions would cause an improvement in the measured VWM performance through the stimulation of the visual processes of the brain VWM relies upon. Participants were between the ages of 18-25 with medically diagnosed ADHD. Experimentation used a pre- and post-test design, where participants took a VWM test as the pre-test, then watched the "spinning dancer" ambiguous illusion for 1 minute, and took the VWM test as a post-test. The results of the VWM test showed the percentage of correct responses and the average reaction time in seconds. The results from the pre-and post-test were compared using paired t-tests. Participants were split into group A, those who reported taking medication for their ADHD, and Group B, those who did not take medication for their ADHD, and data analysis were done separately on these groups. Group A participants exhibited a statistically significant increase in the percentage of correct responses after the illusion was shown, while Group B participants showed no significant differences after the illusion was shown. The results indicated that the viewing of ambiguous illusions is only effective in improving VWM performance in ADHD populations when in combination with prescription ADHD medication.

Keywords: Visual Working Memory, Attention-Deficit Hyper Disorder, Ambiguous Illusions, Visual Perception, Visual Processing

INTRODUCTION

Attention-Deficit Hyper Disorder (ADHD) is a prevalent problem in today's world. In 2016, 10.2% of children, ages four through seventeen, were diagnosed with ADHD in the United States (Xu et al., 2018) and 4.4% of US adults were diagnosed with ADHD (Kessler et al., 2016). ADHD is a neurodevelopmental disorder that can be difficult to live with and can cause many disruptions in the daily lives of those who suffer from it, such as having trouble paying attention, controlling impulsive behaviors, and being overly active (National Center for Disease Control and Prevention [CDC], 2020). Those with ADHD face impairments in different branches of working memory, and deficits in visual working memory (VWM) are significant (Martissun et al., 2005). VWM is a domain of working memory that maintains a limited amount of visual information for a short time so that it can be quickly accessed to serve the needs of ongoing tasks (Luck and Vogel, 2013). VWM is important in several aspects of the lives of humans, such as learning, social ability, and job performance (Johnson, 2016).

Impaired VWM causes difficulties in many aspects of the lives of people who suffer from it such as causing difficulties learning in work and academic settings and an inability to conceptualize problems (Johnson, 2016). VWM is essential for social abilities, therefore those who have deficits in VWM are more likely to have increased awkwardness and decreased social skills overall (Johnson, 2016). VWM is also a measure of overall cognitive ability and general fluid intelligence, and poor VWM is linked to decreased intelligence and cognitive function (Cowan et al., 2005).

Training Visual Working Memory

VWM has shown to be improved through training aspects of the visual systems that it is reliant upon, specifically the visual perception

and visual processing systems of the brain (Luck and Vogel, 2013), by conducting distractor filtering efficiency found by a study in Beijing Key Laboratory of Learning and Cognition Distractor filtering efficiency is the brain's ability to filter out visual distractors (irrelevant visual information) in an image and is therefore dependent on the visual processing system. The study included a 20-day training program on distractor filtering efficiency and improvements from a pre- to post- VWM test were shown (Li et al., 2017). Another study conducted by researchers from the John Hopkins University showed that VWM can also be improved through training in action video games, which have been shown to enhance a range of visual perceptual skills and therefore enhance the visual perception system (Blacker et al., 2014).

Previous research has also shown that VWM can be improved in populations with impaired VWM due to intellectual disabilities. Researchers at the Utrecht University found that the "extensive computerized training programming can effectively improve both VWM and auditory working memory in populations with mild to borderline intellectual disabilities" (Van der Molen et al., 2010).

Ambiguous Illusions

Ambiguous optical illusions are a type of illusion in which images can be seen differently with shifts in perspective. Ambiguous illusions cause the dissociation of "bottom-up to top-down processioning," which causes a shift in the illusion when changing perspectives, and is reversible when shifting back to the original perspective (Wimmer, 2011).

Gap and Current Study

The current research aimed to fill the gaps of the connection between ambiguous illusions and VWM by investigating the research question of "How do ambiguous optical illusions affect the visual working memory performance of populations with attention- deficit hyper disorder (ages 18-25)?" The purpose of this research was to study if the viewing of ambiguous optical illusions could be an effective, practical tool in improving VWM performance in populations with VWM deficits, specifically young adult ADHD populations, and thereby mitigate some of the symptoms associated with it.

It was hypothesized that if ambiguous illusions were viewed, then the performance of VWM in ADHD populations (ages 18-25) would improve significantly through stimulation of the visual processing and perception systems, which VWM is reliant on. It seemed likely that ambiguous illusions would significantly improve VWM through the stimulation of the visual perception and processing systems of the brain due to previous studies which showed that training involving the stimulation of these visual systems led to significant improvements in VWM (Blacker et al., 2014; Klingberg et al., 2010; Li et al., 2017; Van der Molen et al., 2010), suggesting that ambiguous illusions would similarly improve VWM.

METHODOLOGY

This true experimental study involved the collection of quantitative data from the MemTrax memory test, a VWM performance test, from young adult ADHD populations. Data was analyzed with paired t-tests to compare if significant differences were present between the pre- and post-test results on the VWM tests.

Participants

First, test subjects had to be collected. Each subject was between the ages of 18 and 25 and had medically diagnosed ADHD. While ADHD is more commonly associated with children, young adults were used for this research because ADHD is also prevalent among adults, affecting 4.4% of US adults (Kessler et al.,2016). The age group of the subjects, 18-25 known to have the highest VWM performing systems, eliminated the factor of age as a cause of significant impairments observed in VWM performance and made it likely that any observed impairments were due to ADHD.

In order to reach out to a large number of participants with ADHD in the targeted age group, subjects were collected through the ADHD subreddit on the Reddit website. This sampling was used due to limitations of the COVID-19 pandemic disallowing in-person subject collection. A post was made on

the ADHD forum outlining the question, the purpose of the study, and the link to the google form was provided. The post also informed possible subjects that experimentation was completely voluntary and possible risks of viewing the ambiguous illusion included eye strain and headache, and they were encouraged to stop participation if these were to occur.

Method Design

Experimentation followed the template used by many major studies which studied the effect of a variable on VWM performance. This template followed a pre- and post-test design where a pre-VWM test was given to subjects to measure initial VWM performance, then the stimulus occurred, and a post-VWM test was given and the results between the pre- and post-tests were compared. This template has been used to study the effects of distractor filter efficiency training, action video games, learned reward association, and computerized working memory training on VWM performance (Li et al., 2017; Blacker et al., 2014; Gong and Li, 2014; Van der Molen et al., 2010). The method was modeled after these studies due to the shared experimental purpose of studying the effect of a VWM stimulus on VWM performance.

The MemTrax memory test was used to test VWM performance for the pre- and post-test. This is a professional VWM test created by Stanford professor, Dr. Wes Ashford. It was used in experimentation because it is an easily accessible online VWM test and has been professionally tested to be successful in measuring aspects of VWM. A study of 868 subjects completed the test and it was found that test results were consistent with the effects of education and age on memory, proving this test to be an effective memory test. The test showed the test subjects 50 random images, some of which were repeated, and the subjects were prompted to press the spacebar on their keyboard when they believed they saw a repeated image. The test measured the performance of VWM by measuring the accuracy, or the percentage of correct responses, of the subject's VWM and the reaction time, or the average time between when a repeated image was seen and when the spacebar was pressed.

Preliminary Study

Because of financial constraints, only one free format of the MemTrax test was available to be used as both the pre- and post-test in this research. This led to a question in the validity of pre- and post- test research design because taking the MemTrax test once and repeating it again could possibly affect the results of the second test due to familiarization with the test. To ensure the validity of the research design, a small preliminary study was conducted to answer the question of "Is there a significant difference between taking the MemTrax memory test once, and repeating it 1 minute after?". To answer this question, 17 participants between the ages of 18 and 25 were collected through a convenience sample of reaching out to people known by the conductor who met these requirements. These participants did not have ADHD due to difficulties finding participants with this disorder known by the conductor. This experimentation was done virtually due to the COVID-19 pandemic and instructions for this preliminary study were prompted using a google form.

Preliminary Study Methods

The methods included taking the MemTrax test once and recording the results, waiting one minute to account for the minute that would be spent watching the illusion in actual experimentation, and then taking it a second time and recording the results again. Subjects were asked if there were any major distractions that may have affected their abilities on either the pre- or post- test, and if they answered yes their results were not used in the study because the data was affected by outside, uncontrolled variables so it did not give accurate results for the purpose of this study. Out of the 17 subjects collected, 15 responded that no distractions affected their results and their responses were used in the analysis.

Preliminary Study Results

The results showed that the mean

percentage of correct responses in the first test was 93.73%, and in the repeated test it was 94.27%. The mean average reaction time of the first test was 0.79100 seconds and in the repeated test it was 0.75080 seconds. A paired t-test was used to compare the means of the first test and second test percentage of correct responses and average reaction times, and showed that differences in pre-and post-test were statistically insignificant in both, indicating that the difference between taking the MemTrax test once, waiting a minute, and taking it again, is insignificant. This ensured that the pre- and post- test research design was valid and allowed experimentation to begin.

Experimentation

Experimentation was done virtually due to limitations of the COVID-19 pandemic, and subjects were prompted with instructions using a google form. Subjects were prompted to complete the instructions of the study in a low-distraction environment to eliminate confounding variables in their environment which may affect their results. Subjects were first prompted to take the MemTrax memory test as the pre-test and record their percentage of correct responses and average reaction times on the test. The purpose of this pre-test was to get the subjects' baseline VWM performances in ADHD populations of 18-25 without any VWM stimulus prior to taking the test. After taking the pre-test, subjects were prompted to watch an ambiguous illusion, called "spinning dancer illusion," for 1 minute prior to taking the post-test. This illusion depicts a dancer, which can be seen both spinning towards the left and the right with shifts in perspective (Parker-Pope, 2008). The test subjects recorded how many times they are able to see the dancer shift directions while watching the illusion. This number was irrelevant to data collection but ensured that the subjects were focused on the illusion. After watching the illusion, the google form prompted the subjects to take the post-test with the same instructions to take the MemTrax memory test and record their results as they did in the pre-test. Subjects were asked if there were any major distractions that may have affected their abilities on either the pre- or post- test, and if they answered yes, their results were not used in the study because the data was affected by outside, uncontrolled variables.

Recorded results on the MemTrax test from pre- and post- tests were compared with a paired t-test. A t-test was used to see if the difference in the means of the pre- and post- test results of each variable that was tested (the percentage of correct responses and average reaction time) was statistically significant, which would indicate if the effects of viewing ambiguous optical illusions statistically significantly improved the performance of VWM.

DATA AND RESULTS

At the end of data collection, 94 total responses were received. To select if responses were eligible for data collection, they were analyzed to ensure the subject was eligible to be in the study due to being in the targeted population, completing all the tasks, and having no recorded uncontrolled confounding variables affecting their results.

Out of the 94 responses, 76 met all eligibility requirements and were used in data collection. Of these 76 respondents, 54 responded that their ADHD was being treated with prescription medication, while the other 22 responded that they were not being treated with medication. Data analysis was done separately on these two groups as the most common ADHD medications are stimulants which have been shown to improve domains of working memory in adult ADHD populations (Wong and Stevens, 2012; Furmeiar et al., 2016), indicating different baseline VWM performances in the groups. The participants who responded as receiving treatment were labeled as Group A and the participants who responded as not receiving treatment were labeled as Group B.

Group A Results

The data from the MemTrax test that was collected to measure the performance of VWM included the percentage of correct responses and average reaction times before and after viewing the illusion. The results from before and after viewing the illusion were compared in each group.

The mean percentage of correct

responses of Group A in the pretest was 93.63%, while the mean in the post-test was 96.07% (Figure 1) . The mean average reaction time of Group A in the pretest was 0.83913 seconds, while the mean in the post-test was 0.88624 seconds (Figure 2).

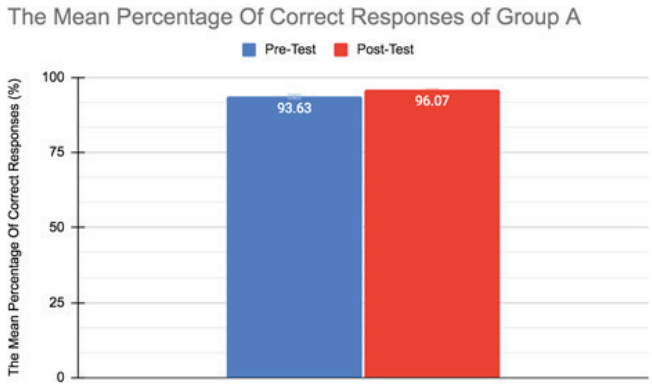


Figure 1: The Mean Percentage of Correct Responses of Group A Subjects in the Pre vs. Post Test

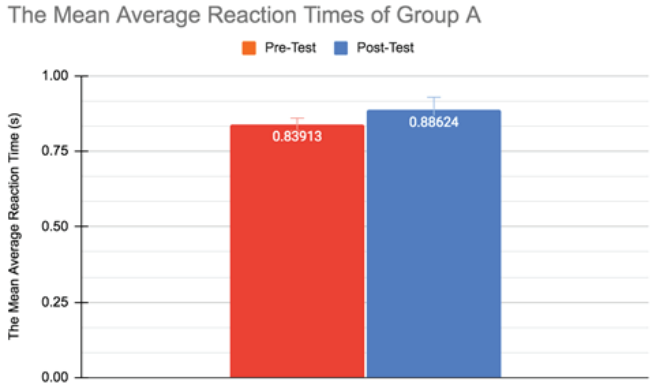


Figure 2: The Mean Average Reaction Time of Group A Subjects in the Pre vs. Post Test

A paired t-test was conducted to test the statistical significance of the difference in the mean percentage of correct response in the pre-and post-test, as well as the difference in the mean average reaction time in the pre- and post-test. When the paired t-test was performed, the t-value was 3.8308. At a 95% confidence interval for 53 degrees of freedom in a two-tailed t-test, the null hypothesis would be rejected when the Group B t-value was greater than 2.0057. Because the t-value was greater than the critical value, the null hypothesis was rejected. The P-value of 0.0003 indicated extremely statistically significant results. Therefore, after the ambiguous illusion was shown to the Group A participants, the mean percentage of correct responses on the MemTrax test significantly increased.

The statistical values of the average reaction times for Group A in the pre- and post-test were also calculated. When the paired t-test was performed to compare the statistical significance of the difference in means of the average reaction times of Group A in the pre- and post-test, the t-value was 2.5559. At a 95% confidence interval for 53 degrees of freedom in a two-tailed t-test, the null hypothesis would be rejected when the t-value was greater than 2.0057. Because the t-value was greater than the critical value, the null hypothesis was rejected. This indicates that the average reaction time significantly increased after the ambiguous illusion was shown to the group A participants. Therefore, when the illusion was viewed, there was a significant increase in the percentage of correct responses, as well as the average reaction time, in Group A participants.

Group B Results

Group B participants reported to not take any medication for their ADHD. The mean percentage of correct responses of Group B in the pretest was 90.96%, while the mean in the post-test was 88.35%, showing a decrease of 2.61% in the mean percentage of correct responses after the illusion was shown (Figure 3). The mean average reaction time of Group B in the pretest was 0.93482 seconds, while the mean in the post-test was 0.88141 seconds, showing a slower mean reaction time of 0.04711 seconds after the illusion was shown (Figure 4).

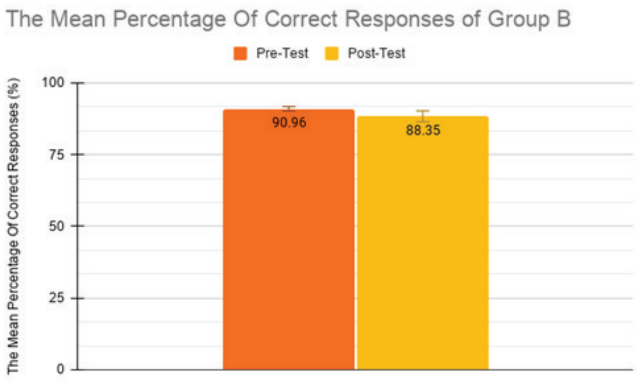


Figure 3: The Mean Percentage of Correct Responses of Group B Subjects in the Pre vs. Post Test

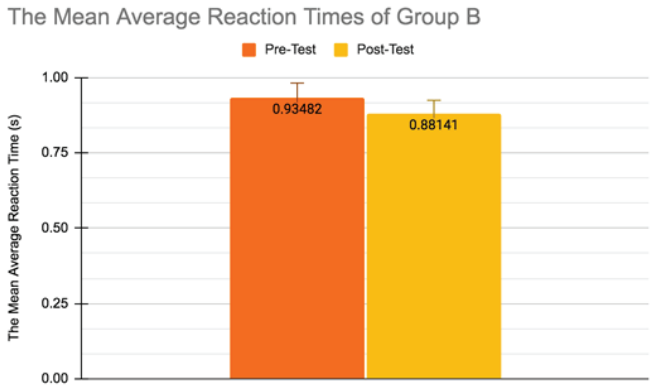


Figure 4: The Average Reaction Time of Group B Subjects in the Pre vs. Post Test

A paired t-test was also conducted to test the statistical significance of the difference in the mean percentage of correct response and average reaction time in the pre-and post-test of Group B participants.

When the paired t-test was performed, the t-value was 1.0736. At a 95% confidence interval for 21 degrees of freedom in a two-tailed test, the null hypothesis would be rejected when the t-value was greater than 2.080. Because the t-value was less than the critical value, the null hypothesis was accepted. Therefore, after the ambiguous illusion was shown to Group B participants, the mean percentage of correct responses on the MemTrax test decreased, but not statistically significantly. This indicates that the illusion had no significant effect on the percentage of correct responses of Group B participants.

The statistical values of the average reaction times for Group B in the pre- and post-test were also calculated.

When the paired t-test was performed to compare the statistical significance of the difference in means of the average reaction times of Group B in the pre- and post-test, the t-value was 0.7184. At a 95% confidence interval for 21 degrees of freedom in a two-tailed t-test, the null hypothesis would be rejected when the t-value was greater than 2.080. Because the t-value was less than the critical value, the null hypothesis was accepted. This indicates that the average reaction time decreased after the ambiguous illusion was shown to Group B participants, but not statistically significantly. Therefore, when the illusion was viewed to Group B participants, there was no statistically significant effect on the percentage of correct responses or average reaction time performance.

DISCUSSION

The results both did and did not support the hypothesis that the viewing of ambiguous illusions would significantly improve VWM performance.

In Group A, both the percentage of correct responses and average reaction time increased significantly in the post-test. The significant increase in the percentage of correct responses in Group A showed a significant improvement in VWM performance after the illusion was shown and reflects an increase of VWM capacity, supporting the hypothesis that ambiguous illusions would improve VWM performance. Multiple studies have shown that working memory capacity has the possibility to be improved through training of the cognitive system (Holmes et al., 2009; Li et al., 2017; Owens et al., 2013), therefore the increase in accuracy when the illusion was shown was most probably caused by the training of the VWM system through the ambiguous illusion. Stimulation of the visual perception and processing systems occurred when the ambiguous illusion was shown (Long and Toppino, 2010), therefore stimulating the same brain mechanisms used by VWM (Luck & Vogel, 2013), leading to an increase in VWM accuracy.

In Group A, the slower reaction time was a significant increase of 0.05 seconds in the post-test. This difference may be a consequence of the research design of repeating the MemTrax test. The preliminary study which tested if there was a significant difference between the first MemTrax test and the repetition of the test found that the difference in both the mean percentage of correct responses and average reaction times was statistically insignificant, but the mean average reaction time did increase by around 0.04 seconds in the post-test. While the difference showed to be insignificant when the paired t-test was conducted, because there were only 15 participants, more trials should be conducted to confirm the difference is insignificant. It is likely that the increase in 0.05 seconds was due to the research design, and the viewing of the illusion did not have a significant effect on the average reaction time,

but more trials of the preliminary study need to be conducted to confirm this.

In Group B, the mean percentage of correct responses and the average reaction time in the post-test decreased, but the difference between the pre- test and post-test showed to be statistically insignificant in both. This shows that the viewing of the ambiguous illusion had no significant effect on the VWM performance of group B participants, which does not support the hypothesis. This could indicate that the viewing of ambiguous illusions is only effective in improving VWM performance in ADHD populations when in combination with prescription ADHD medication which also improves VWM performance, just as behavior therapy for ADHD is most effective when used in combination with stimulant ADHD medication in older children and young adults (Centers for Disease Control and Prevention, 2020).

CONCLUSION

The results showed that the viewing of ambiguous illusions can be used as supplemental training in improving VWM capacity in young adult ADHD populations, but only when in combination with prescription medication. The improvement of VWM performance in this population caused by ambiguous illusions is specific to improvements in VWM accuracy, which was reflected by the significant increase in the percentage of correct responses in Group A subjects, and does not affect the reaction time, as reflected by the inconclusive average reaction time results in Group A subjects.

While the results indicated that the viewing of ambiguous illusions will not affect the VWM performance of young adult ADHD populations who do not take medication, they did show ambiguous illusions could be used as supplemental training to improve VWM performance in young adult ADHD populations when in combination with medication. This information can be utilized by members of this population because this simple task of viewing ambiguous illusions could be used as a practical tool to boost VWM performance, which could lead to improvements in areas affected by VWM, such as job performance, social ability, and learning ability (Johnson, 2016).

REFERENCES

Ashford, J. W., Gere, E., & Bayley, P. J. (2011). Measuring memory in large group settings using a continuous recognition test. Journal of Alzheimer's disease : JAD. https://www.ncbi.nlm.nih.gov/pubmed/21908910.

Blacker, K. J., Curby, K. M., Klobusicky, E., & Chein, J. M. (2014). Effects of action video game training on visual working memory. Journal of Experimental Psychology: Human Perception and Performance, 40(5), 1992–2004. https://doi.org/10.1037/a0037556

Chung, W. (2019, November 1). Prevalence and incidence of adhd among different racial and ethnic groups. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2753787.

Cowan N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav. Brain Sci. 24, 87–114 10.1017/S0140525X01003922

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. Cogn. Psychol. 51, 42–100. doi: 10.1016/j.cogpsych.2004.12.001

Cowan, N., Saults, J. S., & Clark, K. M. (2015, July). Exploring age differences in visual working memory capacity: is there a contribution of memory for configuration? Journal of experimental child psychology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4397185/.

Danielson, Melissa et al. (January 2018). Prevalence of Parent-Reported ADHD Diagnosis and Associated Treatment Among U.S. Children and Adolescents, 2016. Journal of Clinical Child & Adolescent Psychology

Faraone S., Beiderman J., & Mick, E. (2006). The age-dependent decline of attention deficit hyperactivity disorder: A meta-analysis of follow-up studies. Psychological Medicine, 36(2), 159-165. doi:10.1017/S003329170500471X

Fuermaier, A. B. M., Tucha, L., & Koerts, J. (2016, April 18). Effects of methylphenidate on memory functions of adults with ADHD. Taylor & Francis. https://www.tandfonline.com/doi/full/10.1080/23279095.2015.1124108.

Gong, M., & Li, S. (2014). Learned reward association improves visual working memory. Journal of Experimental Psychology: Human Perception and Performance, 40(2), 841-856. doi:10.1037/a0035131

Holmes, J., Gathercole, S. E., and Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. Dev. Sci. 12, 9–15. doi: 10.1111/j.1467-7687.2009.00848.x

Johnson, R. (2016, November 16). Visual Working Memory. https://www.advancedvisiontherapycenter.com/about/blog/e_961/Signs_of_a_Vision_Problem/2016/11/Visual_Working_Memory.htm

Kessler, R. C., Adler, L., Barkley, R., Biederman, J., Conners, C. K., Demler, O., ... Zaslavsky, A. M. (2006, April). The prevalence and correlates of adult ADHD in the United States: results from the National Comorbidity Survey Replication. The American journal of psychiatry. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2859678/.

Klingberg, T., Foressberg, H., & Westerberg, H. (2010, August 10). Training of Working Memory in Children With ADHD. Journal of Clinical and Experimental Neuropsychology. https://www.tandfonline.com/doi/pdf/10.1076/jcen.24.6.781.8395?needAccess=true.

Li C-H, He X, Wang Y-J, Hu Z and Guo C-Y (2017) Visual Working Memory Capacity Can Be Increased by Training on Distractor Filtering Efficiency. Front. Psychol. 8:196. doi: 10.3389/fpsyg.2017.00196

Long, G. M., & Toppino, T. C. (2010). Enduring interest in perceptual ambiguity: Alternating views of reversible figures. Psychological Bulletin, 130, 748–768. doi:10.1037/0033-2909.130.5.748

Luck S. J., Vogel E. K. (1997). The capacity of visual working memory for features and conjunctions. Nature 390, 279–281 10.1038/36846

Luck, S. J., & Vogel, E. K. (2013, August). Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends in cognitive sciences. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3729738/.

Martinussen, R., Hayden, J., Hogg-Johnson, S., and Tannock, R. (2005). A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. J. Am. Acad. Child Adolesc. Psychiatry 44, 377–384. doi: 10.1097/01.chi.0000153228.72591.73

Martel, M. M., von Eye, A., & Nigg, J. (2012, July). Developmental differences in structure of attention-deficit/hyperactivity disorder (ADHD) between childhood and adulthood. International journal of behavioral development. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4307607/.

National Center for Disease Control and Prevention. (2020, September 21). What is ADHD? https://www.cdc.gov/ncbddd/adhd/facts.html.

O'Connor, K. P., Aardema, F., Robillard, S., Guay, S., Pelissier, M.-C., Todorov, C., … Doucet, P. (2006). Cognitive behaviour therapy and medication in the treatment of obsessive-compulsive disorder. Acta Psychiatrica Scandinavica, 113(5), 408–419. https://doi.org/10.1111/j.1600-0447.2006.00767.x

Owens, M., Koster, E. H., and Derakshan, N. (2013). Improving attention control in dysphoria through cognitive training: transfer effects on working memory capacity and filtering efficiency. Psychophysiology 50, 297–307. doi: 10.1111/psyp.12010

Parker-pope, T. (2008, April 28). The Truth About the Spinning Dancer. https://well.blogs.nytimes.com/2008/04/28/the-truth-about-the-spinning-dancer/.

R. Blake, N. K. L., YS. Bonneh, M. P., J. Britz, T. L., HG. Burnett, T. J., A. Caramazza, J. R. S., SM. Carlson, L. J. M., … PD. Zelazo, D. F. (2016 January 1). Assessing perceptual change with an ambiguous figures task: Normative data for 40 standard picture sets. Behavior Research Methods. https://link.springer.com/article/10.3758/s13428-015-0564-5.

Van der Molen, M. J., van Luit, J. E. H., van der Molen, M. W., Klugkist, I., and Jongmans, M. J. (2010). Effectiveness of a computerised working memory training in adolescents with mild to borderline intellectual disabilities. J. Intellect. Disabil. Res. 54, 433–447. doi: 10.1111/j.1365-2788.2010.01285.x

Wimmer, M.; Doherty, M. (2011). "The development of ambiguous figure perception: Vi. conception and perception of ambiguous figures". Monographs of the Society for Research in Child Development. 76 (1): 87–104. doi:10.1111/j.1540-5834.2011.00595.x

Wong, C. G., & Stevens, M. C. (2012, December 29). The effects of stimulant medication on working memory functional connectivity in attention-deficit/hyperactivity disorder. https://www.sciencedirect.com/science/article/abs/pii/S0006322311011267.

Xu, Guifeng et al. (August 2018). Twenty-Year Trends in Diagnosed Attention-Deficit/Hyperactivity Disorder Among US Children and Adolescents, 1997-2016. JAMA Network Open. 2018;1(4):e181471.

An Application of Machine Learning to Predict the Post-Operative Life Expectancy in Lung Cancer Patients

Spandana Pavuluri
duPont Manual High School
10th, Louisville, Kentucky

ABSTRACT

Lung cancer is the third most common type of cancer in the U.S. and thoracic surgery, which can result in death or survival, is a viable option for certain lung cancer patients. This project used Weka machine learning software containing algorithms and attribute selection techniques to determine the most accurate predictive model for post-operative life expectancy in order to increase transparency for patients. Specifically, algorithms, which run through the data and predict whether patients survive or died, were tested to determine which was most accurate. Using Weka, the 23 base algorithms (no attribute selection technique) were run and accuracy percentages were logged. Then, the first selection technique (which essentially removes columns of information that it deems irrelevant to the prediction) was run, the "unselected" attributes were removed, and all 23 algorithms were run again under this attribute selection technique. This process was repeated with 8 other attribute selection techniques resulting in various combinations of attribute selection techniques & algorithms to compare. No combination of algorithm and attribute selection technique got past 85.1064%, which was also reached by the control. A majority of the algorithms didn't have accuracy percentages that were significantly different from the control algorithm, ZeroR, so the null hypothesis was accepted. 6 of the algorithms performed significantly worse than ZeroR, so the alternative hypothesis was accepted in these cases. As for the attribute selection techniques, all but one did not perform significantly differently from when no attribute selection was used, so the null hypothesis was accepted.

INTRODUCTION

According to the CDC, of all cancers, lung cancer is one of the most common types in the United States (2020). Based on the stage and type of cancer, thoracic surgery may be a viable option for lung cancer patients. In general, thoracic surgery usually involves either the chest or the thorax. Thoracic surgery procedures may vary between lung cancer patients based on many factors, including cancer stage, patients' needs, etc.

This project was done using a data set on the Post Operative Life Expectancy in Lung Cancer patients who undergo thoracic surgery. The data set contains information on 470 patients who are non-identifiable. This data set has 17 different attributes, or specific columns of information for each patient, such as their Age, DGN, PRE4, PRE5, etc. The final attribute, also called the class variable, contains information on whether the patient has died or survived within one year of surgery. This is a binary class variable, meaning there are only two options to convey the information: true or false, meaning they died or survived, respectively. Seventy patients in the data set have died within one year of surgery and 400 have survived within one year of surgery.

Machine learning is the use of technology and algorithms that can learn and predict outcomes on unseen data. Weka, a machine learning software that can be used to test different built-in algorithms and attribute selection techniques on a data set, will be used. Attribute selection is when only certain attributes are taken into consideration when predictions and patterns are being made by the algorithms. For example, if a certain attribute is taken into consideration when trying to create an accurate predictive model but this attribute isn't necessarily related to the outcome of the situation, then it is unnecessary to be taken into consideration and could actually cause confusion in the prediction. Another important machine

learning concept is the use of training and testing sets. In a training set, the algorithms can see the class variables (or final prediction) in the sample to learn from them; in a testing set, the algorithm cannot see the class variables so it tries to use what is learned in the training set to predict the class variables of this sample.

Ultimately, the goal for this machine learning project is to create the most accurate predictive model to predict the post-operative survival of a lung cancer patient within one year of surgery. This will be done by testing different algorithms and using attribute selection techniques. A technique called Ten-Fold Cross Validation is used, splitting the data into 9/10 for the training set and 1/10 for the testing set (Vanschoren). The whole process is repeated 10 times, meaning that each tenth of the data will be the testing set once, and then the average will be calculated.

With lung cancer being so common and thoracic surgery having genuine risk, the importance of transparency in the process of deciding whether to undergo a procedure is essential. Although it may not always be possible to provide patients with a confirmed prediction of what may happen, it must be ensured that any prediction provided to the patient is as accurate as possible. As aforementioned, thoracic surgery can be a life-or-death operation and although there may not be many alternatives to take to treat lung cancer, it is essential that patients feel like they've been given the necessary information to make an informed decision.

The purpose of this project is to increase transparency in the process of a patient's decision on whether they will undergo thoracic surgery to treat their lung cancer or not. While this model may not end up being a hundred percent accurate, the goal of this project is to make it as accurate as possible so that patients can weigh the risks and benefits of the procedure. The question for this project is "Which algorithm and attributes provide the most accurate prediction of survival or death within one year for postoperative lung cancer patients?"

The hypothesis is "If the logistic regression algorithm with a specific feature selection technique is used, the predictive model will be the most accurate." This is determined because logistic regression is generally a highly suitable algorithm for many binary classification problems. It is a bit more difficult to determine which feature selection technique will create the most accurate prediction prior to testing so instead, it was included that using an attribute selection technique will be a factor in increasing the accuracy of the model.

This project differs from past studies because different algorithms and attribute selection techniques will be used to produce the most accurate prediction. It does not use the boosting option, which has been used in other studies, and focuses on whether patients will survive or die rather than how long patients will survive. Essentially, this project uses machine learning to get the most accurate prediction of postoperative life expectancy but deals with different aspects of lung cancer survival and the approach tunes different parameters to determine which is the most accurate prediction.

METHODOLOGY

The Thoracic Surgery dataset was first downloaded and then opened in Weka. Then, to ensure there were no missing values in the data set, the different attributes were clicked on and the missing values were all equal to 0%. To begin experimentation, under the Classify window, the algorithms were selected to be run. 23 algorithms were chosen, either randomly or selectively due to high base accuracy percentages, from all the algorithms available in Weka. The first algorithm, NaiveBayes, was chosen, cross validation with 10 folds was selected, and the algorithm was run. This process was repeated with the other 22 algorithms, which were Logistic, SGD, SMO, IBK, J48, RandomForest, RandomTree, REPTree, VotedPerceptron, Bagging, NaiveBayesMultinomialText, SGDText, AttributeSelected Classifier, CVParameterSelection, MultiScheme, RandomSubSpace, Stacking, Vote, WeightedInstancesHandlerWrapper, InputMappedClassifier, HoeffdingTree, and ZeroR, the control algorithm.

Since the base algorithms had been run, attribute selection techniques were run next. Nine attribute selection techniques were randomly selected and each technique was used with each algorithm, resulting in 230 total combinations to examine, including the 23 base algorithms performance. First, the CfsSubsetEval

attribute selection technique was chosen with the "GreedyStepwise" search method. Once the other attributes were removed, the same process of running all 23 of the algorithms was done with only those selected attributes.

Once all the algorithms had been run under the CfsSubsetEval attribute selection technique, the process was repeated with CorrelationAttributeEval, ReliefAttributeEval, SymmetricalUncertAttributeEval, and InfoGainAttributeEval, which all used the Ranker search method. With the next attribute evaluators, the search method was ranker meaning that instead of saying explicitly which attributes should be selected, the attributes are ranked. For CorrelationAttributeEval, all values ranked above 0.04 were considered as selected values and then this attribute selection technique was run again with values above 0.05 considered as selected. For ReliefAttributeEval, all values ranked above 0.01 were considered as selected values and this technique was re-run with values above 0.02 considered as selected. For SystemUncertAttributeEval, all values ranked above 0 were considered as selected values and then this attribute selection technique was re-run with values above 0.004 considered as selected. InfoGainAttributeEval was first run with values above 0 being considered as selected and then with values above 0.002 considered as selected. Since CorrelationAttributeEval, ReliefAttributeEval, SymmetricalUncertAttributeEval, and InfoGainAttribute Eval were each run twice with different selected values, there were a total of 10 different attribute selection techniques (including the base with no selection technique) that were run in this project. After all of the combinations of algorithms and attribute evaluators were run, the accuracies of all the different combinations were compared by sorting through the data that was saved.

DATA AND RESULTS

Accuracy percentages for each base algorithm and each algorithm and attribute selection technique combination were recorded and accuracy values containing the highest accuracy (85.1064%) were highlighted. No accuracy percentages got above the 85.1064% which was also reached by the control algorithm, ZeroR. The majority of the algorithms that had the highest accuracy percentage of 84.1064% with base, meaning no attribute selection techniques, tended to have the most highlighted cells total among different attribute selection techniques. The algorithms that had been added during the project proved to be more effective than the 11 first randomly selected algorithms, as can be seen with the mean values listed in Table 1. Another trend that has been apparent from the raw data is that the accuracy values among the last 12 algorithms were all the exact same, even among different attribute selection techniques. Additionally, the raw data shows that certain attribute selection techniques, specifically CfsSubsetEval, may not have been as effective.

To answer the project's question, there were numerous different combinations of algorithms and attribute selection techniques that were able to perform the most accurately.

Algorithm 1	Mean	df	Calculated *t*-value	Critical *t*-value
NaiveBayes	81.36169	18	-6.005968657	2.101
Logistic	82.61702	18	-6.900896724	2.101
SGD	84.48936	18	-0.8309802073	2.101
SMO	84.65956	18	-0.3932520207	2.101
IBK	79.9149	18	-5.130407451	2.101
J48	84.4681	18	-0.7333781718	2.101
RandomForest	84.51065	18	-3.772199604	2.101
RandomTree	78.34042	18	-9.083717546	2.101
REPTree	84.383	18	-1.083241897	2.101
VotedPerceptron	84.48937	18	-0.7385327644	2.101
Bagging	84.02128	18	-2.146995251	2.101
NaiveBayesMult Text	84.80853	18	0	2.101
SGDText	84.80853	18	0	2.101
Att Select Classifier	84.80853	18	0	2.101
CVParameter	84.80853	18	0	2.101
MultiScheme	84.80853	18	0	2.101
RandomSubSpace	84.80853	18	0	2.101
Stacking	84.80853	18	0	2.101
Vote	84.80853	18	0	2.101
WeightedIns.	84.80853	18	0	2.101
InputMClassifier	84.80853	18	0	2.101
HoeffdingTree	84.80853	18	0	2.101
ZeroR (control)	84.80853	18	0	2.101

Table 1. T-test Comparing 23 Algorithms and Control Algorithm, ZeroR

To determine whether any algorithm was statistically significantly different from the performance of the control algorithm, Zero R, a t-test was performed. Each algorithm's data was compared to ZeroR and the values for every t-test performed were either negative or equal to 0, as shown in the table. Seeing as the critical t-value was 2.101, all of the calculated t-values that were between -2.101 and 2.101 meant those algorithms had no significant statistical difference between their accuracy percentages and the ZeroR (no intelligence) algorithm accuracy percentages. This means that SGD, SMO, J48, REPTree, VotedPerceptron,

NaiveBayesMultinomialText and all the algorithms following it supported the null hypothesis. Since the t-test performed was two-tailed, if the t-value was out of the range of -2.101 and 2.101, then the null hypothesis would be rejected. The NaiveBayes, Logistic, IBK, RandomForest, RandomTree, and Bagging algorithms support the alternative hypothesis meaning that these algorithms performed significantly worse than the ZeroR algorithm.

Attribute Selection Technique	Mean	df	Calculated *t*-value	Critical *t*-value
CfsSubsetEval	82.11843913	44	-2.68038357	2.042
CorrelationAttributeEval	83.70953043	44	-0.00001083195165	2.042
ReliefAttributeEval	83.98706087	44	0.3639777951	2.042
Sym.UncertAttrib.Eval	84.23682609	44	0.7874720459	2.042
InfoGainAttributeEval0	84.24607826	44	0.8103036636	2.042
CorrelationAttributeEval.05	84.36634348	44	0.9946989099	2.042
ReliefAttributeEval.02	83.6817913	44	-0.03184462204	2.042
SymmetricalUncertEval.004	84.26457391	44	0.8263484678	2.042
InfoGainAttributeEval.002	84.26457391	44	0.8263484678	2.042
Base	83.70953913	44	0	2.042

Table 2. T-test Comparing 23 Algorithms and Control Algorithm, ZeroR

The t values for the attribute selection techniques when compared against their control (base- no attribute selection technique) range only between -1 and 0.5. Seeing as most of these values are all within the range of -2.042 and 2.042, mostly all of the different attribute selection techniques did not prove to have a statistically significant difference in results compared to when no attribute selection technique was used on the algorithms. This means that for all but one of the various attribute selection techniques, the null hypothesis was accepted. The CfsSubsetEval attribute selection technique is the only attribute selection technique that has a t-value out of the range of -2.042 and 2.042 and supports the alternative hypothesis meaning it performed significantly worse than when no attribute selection technique was used. When comparing the means of the different attribute selection techniques, CorrelationAttributeEval, with values above .05 considered as selected, had the highest accuracy percentages.

CONCLUSION

The data showed that the base algorithms and the combinations of algorithms and attribute selection techniques weren't able to surpass the accuracy percentage of 85.1064%. There were numerous algorithms that performed the "best," at this accuracy percentage of 85.1064%, including the control algorithm, ZeroR. This algorithm uses no intelligence and predicts the majority outcome (University of Waikato, 2020). The techniques were likely not able to determine clear patterns among the data set in order to create accurate predictive models above 85.1064%. More specifically, since this data set had only 70 cases of patients who had died within one year of the surgery vs. 400 patients that had survived, the algorithms likely weren't able to properly distinguish between both outcomes since there was not as much information on patients who had died.

The data also showed that the last 12 algorithms run, which were intentionally selected since they had the highest base accuracy percentages, were consistently more effective than the first 11 algorithms. Therefore, it can be concluded that selective choosing of base algorithms for this data set can cause accuracy percentages to be consistently higher than if algorithms were randomly selected.

Additionally, the last 12 algorithms all had the exact same values, even when they had been run under different attribute selection techniques. This was likely because they had all started with the same value of 85.1064% so the way the algorithms found patterns in the data and predicted was fairly similar. Therefore, even with different attributes removed, they got the same accuracy prediction values since they used very similar methods.

When t-tests were performed on the data, it was found that a majority of the algorithms did not have accuracy percentages that had significant statistical differences from the control algorithm. 6 algorithms performed significantly statistically worse than the control algorithm. These select few algorithms supported the alternative hypothesis stating that " If various algorithms/attribute selection techniques are used, the predictive accuracy will decrease compared to that of the control." When t-tests were performed on the attribute selection techniques, all but one of the

attribute selection techniques had a significant statistical difference compared to when no attribute selection technique was used. Therefore, for a majority of the algorithms and attribute selection techniques, the null hypothesis stating that "If various algorithms and attribute selection techniques are used, the predictive accuracy will not differ significantly from the predictive accuracy of the control," was accepted.

Another study ran a couple of algorithms on the thoracic surgery data set, some of which were also used in this project, and calculated accuracy percentages along with other factors (Khalaf & Wang, 2018). The algorithms, similar to many in this study, were NaiveBayes, Logistic Regression, RandomForest, SVM, and RBFN. Naive Bayes had the worst accuracy in the mentioned study and this project also showed that Naive Bayes was one of the worst predictive algorithms. The similarities in the results of both projects add increased credibility to the findings.

Another study worked to test a couple of algorithms, such as Logistic Regression, J48, Naive Bayes, and determine whether they were accurate predictions. The main differences were that the age attribute was discretized, meaning the values were sorted into different "bins" which were named with a value. Discretization is an important step for data sets with discrete data similar to the thoracic surgery data set (Lustgarten, Gopalakrishnan, Grover, & Visweswaran, 2008). Therefore, the way that this study discretized the age attribute was likely a reason that some of the algorithms performed better with those methods versus this project's methods (the data was not discretized). Additionally, another difference was that the first 5 attributes were removed since they weren't relevant to the prediction but that practice was not done in this project.

For future research, more attribute selection techniques could be run on the algorithms in order to see whether accuracy percentages could improve. Additionally, while all the base algorithms had been run and the ones with the best predictions had been chosen and included in the project, this project could be expanded in the future by running all of the algorithms, even the ones which had not performed the "best" without any attribute selection techniques, to get an even larger range of accuracy percentages. Data discretization could be used in another version of this project to see whether it is able to consistently increase the accuracies of the algorithms. It is important to continue to try to increase the transparency for patients and by running algorithms and attribute selection techniques among other postoperative or other health-related data sets, more accurate predictive models can be extremely beneficial to patients.

REFERENCES

Bartholomai, J. A., & Frieboes, H. B. (2018). Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). doi:10.1109/isspit.2018.8642753

Binary classification with logistic regression. (2017). https://gluon.mxnet.io/chapter02_supervised-learning/logistic-regression-gluon.html

Desuky, A. S., & Bakrawy, L. M. E. (2016). Improved Prediction of Post-operative Life Expectancy after Thoracic Surgery. Advances in Systems Science and Applications, 16(2), 70-80. https://ijassa.ipu.ru/index.php/ijassa/article/view/351

Iduye, S., & Zhuang, X. (2018). Effectiveness of Data Mining Algorithm in Predicting Post Lung Resection Surgery Prognosis, 13. Retrieved from http://cjni.net/journal/?p=5694

Khalaf, A. F., & Wang, Y. (2018). Predicting Post-Operative Survivability of Lung Cancer Patients: A Data Mining Approach. Predicting Post-Operative Survivability of Lung Cancer Patients: A Data Mining Approach. Retrieved from https://www.binghamton.edu/centers/seorl/docs/c018.pdf

Lung Cancer Statistics. (2020, September 22). https://www.cdc.gov/cancer/lun (CDC, 2020)g/statistics/index.htm

Lustgarten, J., Gopalakrishnan, V., Grover, H., & Visweswaran, S. (2008, November 6). Improving classification performance with discretization on biomedical datasets. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656082/

Vanschoren, J. (n.d.). 10-fold Crossvalidation.https://www.openml.org/a/estimation-procedures/7

ZeroR (weka-dev 3.9.5 api). (2020, December 21). Retrieved from https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/ZeroR.html

The Classification of EMG Signals using Machine Learning for the Construction of a Silent Speech Interface

Varun Chandrashekhar
duPont Manual High School
11th, Louisville, Kentucky

ABSTRACT

With 7.5 million people suffering from speech impediments, it is imperative that accurate speech aids are developed. Conditions such as stroke, ALS, and cerebral palsy leave their patients unable to speak and force them to use cumbersome and inefficient devices such as eye or cheek trackers. In this study, a speech aid known as a Silent Speech Interface (SSI) was created. This device could be used by patients with speech disorders to communicate letters in the English alphabet voicelessly, merely by articulating words or sentences in the mouth without producing any sounds. This device captures and records the subtle neurological activation of the muscles in the internal human speech system from the surface of the skin. These EMG signals are then classified into speech in real-time using a trained Machine Learning (ML) model. This device could effectively determine what was communicated with 80.1% accuracy. For this device, it was found that the Support Vector Machine algorithm was the most effective ML model for the classification of EMG signals from the throat. These findings fill the lack of research on optimal ML models for use in an SSI. Overall, this study involves the creation of a device that measures biomedical signals and translates them into speech using the SVM model with high accuracy. This study's findings could improve the accuracy of future SSIs by showing which algorithms are most accurate in practice.

Keywords: Silent Speech Interface, sEMG, Machine Learning, Convolutional Neural Network, Pattern Recognition

INTRODUCTION

Multiple Sclerosis (MS) is a progressive neurodegenerative disease characterized by lesions in the nervous system that affects nearly 2.3 million people worldwide (Brown, 2000). As the disease progresses, MS creates communication problems between the brain and body. Two major impairments that come with MS are speech disorders known as dysphonia and dysarthria (Brown, 2000). Dysphonia affects speech muscles, which can lead to patients being inaudible (Beukelman and Garrett, 1988). Other diseases such as Motor Neuron Diseases (MNDs) cause patients' speech to become unclear, taking away a patients' ability to speak. MND patients are forced to use eye or cheek tracking speech aids which make the user perform specific muscle movements to select letters and words the user wants to communicate. These cumbersome devices prove to be an extremely slow and fatiguing solution for communication. In this study, these systems will be referred to as Conventional Speech Interfaces (CSI).).

I. Silent Speech Interfaces and the Electromyograph (EMG) Signal

Silent speech refers to the act of minimally/internally articulating words without producing sounds (Kapur, 2019) Although silent speech is inaudible, it produces signals that can be recorded and classified into words using Machine Learning. These signals are ElectroMyoGraph (EMG) signals which are created by subtle muscle contractions. During silent speech, speech muscles (cheek, lips, etc.) contract, producing EMG signals in certain patterns. When the same words are spoken, the same muscle contractions occur to produce specific EMG signal patterns. Thus, if the EMG signals can be recorded, it is possible to translate the signal patterns to determine the speech that was silently spoken. This allows for

development of a new type of speech interface known as Silent Speech Interfaces (SSI). SSIs are a more effective speech aid compared to CSIs. However, an SSI's accuracy is highly dependent on the computer algorithms that are used to translate the EMG signals into speech (Kapur, 2019).

The most common method to record EMG signals involves placing electrodes on the skin to detect muscle contractions (Kapur. 2019).

II. Artificial Intelligence for Construction of Automated Silent Speech Interface

SSIs make use of Machine Learning to identify sEMG/EMG patterns to translate silent speech into language (Denby, 2011). ML is a broad field and can be subdivided into 2 main categories: supervised learning and unsupervised learning. The construction of an SSI requires supervised learning, in which an algorithm generates a function that maps inputs to desired outputs (Ayodele, 2010).

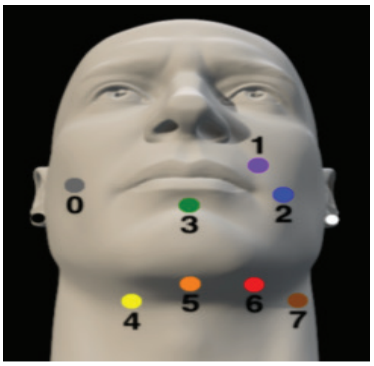
Using supervised learning, it is possible to "teach" and develop an ML algorithm that can translate sEMG/EMG signals into the letters/words that were silently spoken.

In this study, sEMG signals will be translated into one of the five vowels. Thus, ML classification algorithms were used to classify sEMG signals into the letters/words that were silently spoken.

In the case of EMG signal classification, only several supervised ML algorithms could be used for the construction of an SSI, because only a few ML algorithms are capable of analyzing and processing signals (series of numbers). This narrowed the ML models that could be used and set the scope of this study. Access to ML Development Software has allowed several researchers to classify EMG signals.

III. Previous Findings

EMGs are typically recorded through an electromyograph. These high-end machines are expensive and thus an inconvenient solution to monitor EMG signals for the development of an SSI. A study conducted by Kareem et al. used the Myoware to record EMG signals (2017). By comparing the sEMG signals recorded by the Myoware sensor to those recorded by the Electromyograph, Kareem et al. found that the Myoware can be used in ML applications due to its high accuracy (2017). Furthermore, it was determined that sEMG signals recorded from the Myoware and the electromyograph have the same patterns. Therefore Kareem et al. identified that classification of sEMG signals is possible using the Myoware due to its high accuracy. This study was important as it justified the use of a Myoware sensor (used in this study) as an alternative to the electromyograph used in other studies. A landmark paper, titled "Non-Invasive Silent Speech Recognition in Multiple Sclerosis with Dysphonia," by Kapur et al. is one of the most advanced research on the implementation of an SSI. The SSI created, recorded EMG signals from a multitude of locations from the face and throat as shown in Figure 1. These signals were used to train, validate, and test the Convolutional Neural Network (CNN) --an ML algorithm-- that was used to build the SSI. Although the use of the model was never justified, the CNN model yielded high accuracy of 79%. The SSI developed improved the speed/accuracy of communication compared to CSIs (Kapur, 2019). This study was crucial as it laid the foundation for the methodology in this research as it was the only study that identified steps to create an SSI



.

Figure 1: Picture of Electrode Locations - This figure shows an image of electrode placements used in Kapur's study

Another study by Shultz et al. developed an SSI using the EMG-PIT corpus, a database of EMG recordings from the speech system. Using a Gaussian Mixture Model (GMM), the developed SSI managed to achieve a low error rate of 10%. Similar to Kapur et al.'s study, Shultz's study lacks justification for the ML algorithm used. This lack of justification could suggest that other ML algorithms would perform better at classifying sEMG signals in an SSI.

IV. Research Goals & Research Gap

This project aims to construct two types of ML algorithms in order to gain insight into the question: "Which type of Machine Learning Algorithm (Convolutional Neural Networks or Pattern Recognition) is most accurate at classifying surface ElectroMyoGraph (sEMG) signals from the submental triangle (area under the chin) to develop a Silent Speech Interface?" This study would help further improve the ability of SSIs to translate sEMG signals into speech, allowing for more accurate communication. The engineering goal of achieving an 80% accuracy was developed as other studies on SSIs also strived to acquire an 80% accuracy. The engineering goal also involved creating an SSI using a low-cost muscle sensor (Myoware) as previous studies only used electromyographs (Karlik, 2018).

METHODOLOGY

To identify the most accurate model through the true quantitative experimental method, the tested models' F1 scores were compared just as in Karlik's study. F1 scores range from 0 to 1, and high scores indicate that a model is both accurate and precise (Eremenko et al., n.d.). The models' F1 scores were compared, and the model with the highest score was identified as the best performing ML algorithm, answering the research question.

To determine if the created SSI met the engineering goal only the accuracy of the ML model with the highest F1 score was considered. This is because the SSI will use the best performing model (model with highest F1 score - identified in experimental method) to translate EMG signals.

To evaluate models, the SSI had to be created. Creating an SSI involved the following procedures which were also carried out in Kapur's study:

I. Arduino-Based EMG Recorder Development

The first step in the development of the EMG recorder was wiring components together. The EMG recorder would make use of a microcontroller (Arduino) that takes EMG recordings from a sensor and saves data to an SD card. The recorder has buttons to start recording EMG signals.

For the Arduino to record EMG signals, the Myoware muscle sensor was used to detect EMG signals. As discussed previously, the MyoWare is the best commercially available muscle sensor and has been used in other studies due to its reliability and accuracy (Hartman, n.d.). Although Kapur does not use the Myoware, Kareem justifies the use of this sensor as it produces accurate results (n.d.).

The Arduino Mega, Myoware, and button were all connected as shown in the schematic (Figure 2) and diagram (Figure 3) below. To ensure functionality, the device was connected this way according to data sheets provided by the Arduino company (n.d.).

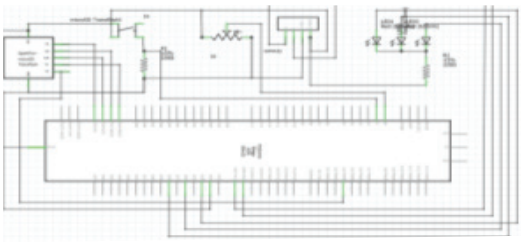


Figure 2: Schematic of EMG Recorder - This figure shows the EMG recorder schematic

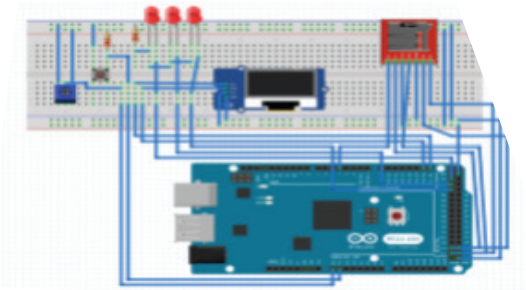


Figure 3: Diagram of EMG recorder - This figure shows the EMG recorder wiring diagram

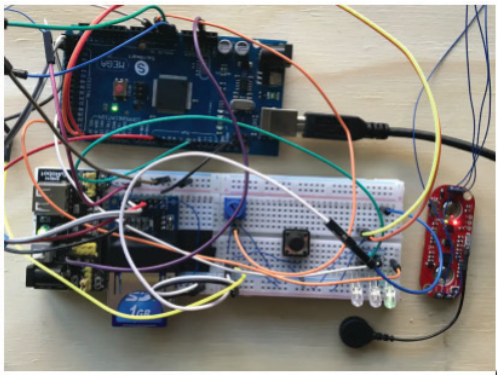


Figure 4: Image of Developed EMG Recorder - This figure shows an image of the created EMG recorder with the red Myoware sensor to the right and the blue Arduino Mega at the top

An image of the fully constructed EMG recorder is shown in Figure 4. This EMG recorder was programmed to perform various tasks. The device has 3 tasks.

Tasks 1 and 2 involve taking in user input. The code for the button allows the device to start recording EMG data only when the user is ready to speak. This was also done by Kapur and ensures that EMG data is only recorded when silent speech is produced. The code to determine what letter is being silently spoken is important as the device needs to associate each EMG recording with a specific letter. After EMG data has been recorded, the EMG data has to be saved.

II. Database Preparation

A database needed to be created to train the ML models. This data was created using the created EMG recorder device (Figure 4). Three electrodes were attached to the submental triangle, the area under the chin. A total of 1020 EMG recordings were taken, 170 for each of the 5 vowels and another 170 to establish a baseline of not speaking at all.

Afterward, the EMG dataset was parsed into 2 sets: a training set, with 80% of the entire data, and a testing set, with 20% of the entire data, which was split this way to ensure the magnitude of the training data was sufficient (Mwebaze & Owomugisha, 2016)

III. Pattern Recognition Methods

Once imported into the MATLAB programming environment, the Classification Learner App was utilized to develop the Pattern Recognition Algorithms to classify sEMG signals (Eremenko et al., n.d.). Only 7 types of Pattern Recognition algorithms were capable of translating EMG signals and they were all implemented: Support Vector Machine (SVM), Ensemble, K-Nearest Neighbors (K-NN), Decision Tree Classification, Naive Bayes (NB), Linear Discriminant, Quadratic Discriminant. The EMG data and the corresponding vowel/letter were imported into the computer. Each algorithm was trained/tested on the same dataset to ensure the validity of the results. Due to time constraints and limited computing power, each algorithm was given only 10 iterations (opportunities) to learn from the data, ensuring that no ML model had an advantage over the other tested algorithms. After training each PR model, each algorithm was tested on the previously developed testing data to determine classification accuracy and F1 scores.

IV. Convolutional Neural Network Methods

The GoogleNet algorithm was repurposed for this study to classify EMG signals into letters. This process of repurposing classification layers (Figure 5) from an existing model is known as transfer learning which is beneficial as the CNN model developed using GoogleNet will likely have higher accuracy than other CNN algorithms (MathWorks, Classification Learner App n.d.).

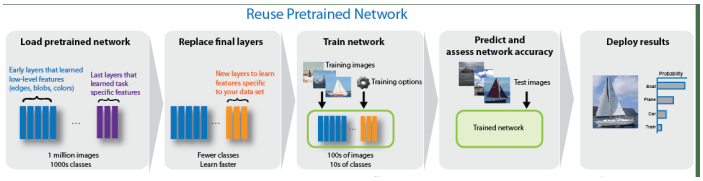


Figure 5: Transfer Learning Implementation - This figure shows the steps required to use transfer learning for an application

Because CNN's require image inputs for classification/training, the signal which was originally a series of numbers had to be converted into an image by converting it to a spectrogram which is a visual representation of signals and was also used in Kapur's study when developing an SSI. A spectrogram for each of the 1020 EMG signals in the dataset was built.

V. Creation of Silent Speech Interface

An SSI is a speech aid that records silent speech and uses an ML algorithm to translate the recorded EMG signals. Therefore the developed SSI has to be able to record EMG signals and then translate those signals using ML which was done by combining the previously built EMG recorder and a computer running ML algorithms.

DATA AND RESULTS

As discussed in the methodology, the ML algorithms were tested using the same testing set. The classification accuracy and F1 scores were calculated for each tested model. F1 scores are

commonly used by data scientists to compare ML models and determine which model is holistically better (Wood, 2019). The classification accuracies and F1 scores for the PR and CNN models are shown in Table I.

MODEL	CLASSIFICATION ACCURACY	F1 SCORES
CONVOLUTIONAL NEURAL NETWORK (CNN) - GOOGLENET	54.90%	0.60
SUPPORT VECTOR MACHINE (SVM) - GAUSSIAN	80.10%	0.81
ENSEMBLE - BAGGED TREES	74.60%	0.73
K-NEAREST NEIGHBORS (KNN) - WEIGHTED	66.70%	0.70
TREE - MEDIUM	59.80%	0.65
NAIVE BAYES - KERNEL	59.30%	0.65
QUADRATIC DISCRIMINANT	55.50%	0.54
LINEAR DISCRIMINANT	49.50%	0.48

Table I: Classification Accuracy & F Scores

As seen in Table I, the SVM model achieved the highest classification accuracy and F1 score. The accuracy values of the ML models tested ranged from 49.5% to 80.1% while the F1 scores ranged from 0.48 to 0.81. It can also be seen that the F1 scores closely correlated with the classification accuracy for each model and were often only ± 0.02 away from the classification accuracy.

Confusion Matrices are a common way to depict the accuracies of ML models. The Confusion Matrices are shown for the SVM model (Figure 15), which had the highest F1-score/accuracy, and the CNN model (Figure 16) which was the only non-PR algorithm tested.

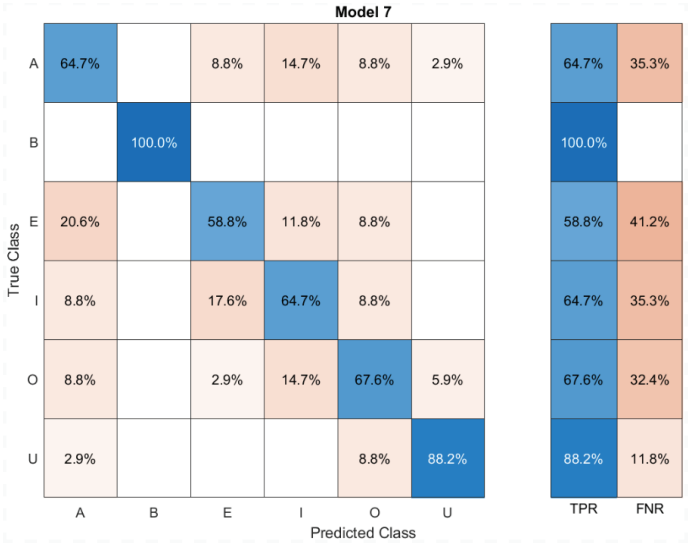


Figure 6: Confusion Matrix of Support Vector Machine which has an 80.1% accuracy and F1 score of 0.81

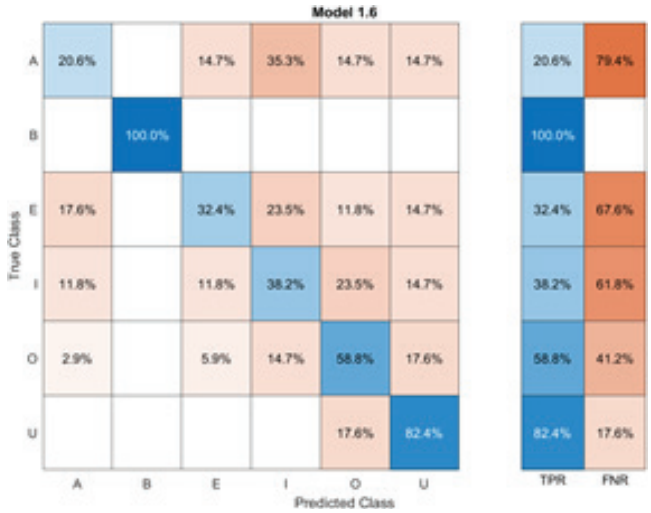


Figure 7: Confusion Matrix of the CNN which has a 54.9% accuracy and F1 score of 0.60

In Confusion Matrices (Figures 6 & 7), the rows denote what letter was silently spoken, whereas the column shows which letter was predicted by the trained ML algorithm. Therefore, all the correct predictions lie along the diagonal vector shaded in blue whereas incorrect predictions are shaded in orange.

Through analysis of the Confusion Matrix, it can be seen that both the SVM model (Figure 6) and CNN model (Figure 7) classified the signals for "not speaking" denoted by "B" (Blank) on the axes, with a 100% accuracy. The SVM model's largest error was due to misclassifying the EMG signals (for letter "E") as the letter "A". This accounted for 20.6% of incorrect predictions associated with the letter E. The CNN model's

largest error was due to misclassifying the EMG signals (for letter "A") as the letter "I". This accounted for 35.3% of incorrect predictions associated with the letter "A".

Both the SVM and CNN models performed poorly when classifying the letters "A" and "I". This misclassification trend also occurred in the other tested PR models indicating that the EMG signals for "A" and "I" are hard to decipher. The SVM model had accuracies of 64.7% and 56.8% for the letters "A" and "I" respectively whereas the CNN model had accuracies of 20.6% and 32.4% for the same letters. These accuracy values for individual letters were the lowest and brought down the overall F1-score/ accuracy.

CONCLUSION

I. New Understanding & Conclusions

An SSI was created and 8 ML algorithms were tested. The calculated F1 scores of each model were used to determine the best performing model and the classification accuracy of the model with the highest F1 score was used to determine whether the engineering goal was achieved

It was found that the highest F1-score (0.81) was achieved using the SVM model which is a type of PR algorithm. This means that the SVM Pattern Recognition model is the most accurate ML algorithm to use in an SSI for classifying EMG signals into speech, thus answering the research question. This is the first study to have ever identified the most accurate ML algorithm for classifying EMG signals in an SSI.

The classification accuracy for the SVM model, the best performing model, is 80.1%. This classification accuracy meets the engineering goal of 80% accuracy.

II. Explanation of Findings & Other Research

The SVM model likely outperformed the CNN model as it is a simpler model and can train on data quickly. Because the number of iterations each model could train for was limited, the CNN wasn't able to fully train in the given amount of iterations. Therefore the SVM algorithm performed better.

The PR models tested had a wide range of accuracies. This can be attributed to the fact that some models, such as the Linear Discriminant, divide numerous classes ineffectively with simple methods. Although these algorithms perform well with small inputs, it wasn't useful to classify large EMG signals.

Karlik found in his study that the CNN is the most accurate algorithm for classifying EMG data for arm prostheses. This study had different results because the nature of EMG signals from the arm and throat are different resulting in different optimal ML algorithms. Additionally, in this study, time and resource limitations could have impacted the performance of tested ML algorithms.

V. Limitations, Future Research & Implications

One limitation of this study is that it used a small dataset. Due to time constraints, the dataset developed to train/test the ML model contained only 1020 signals. With more training data, the accuracy of the ML algorithms would improve. Another limitation is that this study cannot definitively identify the SVM algorithm as the most accurate ML model. There are many types of ML algorithms, such as Artificial Neural Networks, that were not tested in this study due to time constraints and the complexity of models. Future studies can address these limitations by evaluating/comparing more types of ML algorithms. Additionally, future studies could also develop datasets for a multitude of words from the English language. This would allow for the development of a more complete SSI that can truly be used in the real world.

This study's findings can help improve the accuracy of future SSIs by showing that SSIs can achieve better accuracy by using an SVM model. Additionally, the findings of this study can push researchers to develop SSIs without the use of electromyographs as this study was able to achieve good results using a low-cost muscle sensor.

REFERENCES

Alaskar, H. (2018, December). Convolutional neural network application in biomedical signals. Retrieved November 24, 2020, from https://pdfs.semanticscholar.org/fd36/d9c6abf405af2c12d1596162421d3f503591.pdf

Arduino. (n.d.). Arduino Mega 2560 rev3. Retrieved December 18, 2020, from https://store.arduino.cc/usa/mega-2560-r3

Arslan, Y. Z. (2006, July). Observations on the characteristics of EMG signals recorded at different depths. Retrieved November 24, 2020, from https://www.researchgate.net/figure/Penetration-depth-of-needle-electrode_fig1_241204405

Retrieved November 24, 2020, from https://www.researchgate.net/publication/221907660_Types_of_Machine_Learning_Algorithms

Beukelman, D., & Garrett, K. (1988, July 12). Augmentative and alternative communication for adults with acquired severe communication disorders. Retrieved November 24, 2020, from https://www.tandfonline.com/doi/abs/10.1080/07434618812331274687

Ayodele, T. (2010, February). Types of machine learning algorithms. Retrieved November 24, 2020, from https://www.researchgate.net/publication/221907660_Types_of_Machine_Learning_Algorithms

Beukelman, D., & Garrett, K. (1988, July 12). Augmentative and alternative communication for adults with acquired severe communication disorders. Retrieved November 24, 2020, from https://www.tandfonline.com/doi/abs/10.1080/07434618812331274687

Bishop, C., & Nasrabadi, N. (1970, January 01). Pattern recognition and machine learning: Semantic scholar. Retrieved November 25, 2020, from https://www.semanticscholar.org/paper/Pattern-Recognition-and-Machine-Learning-Bishop-Nasrabadi/668b1277fbece28c4841eeab1c97e4ebd0079700

Bonaccorso, G. (2017, July 24). Machine learning algorithm. Retrieved November 25, 2020, from https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781785889622

Brown, S. (2000, October 01). Swallowing and speaking challenges for the MS patient. Retrieved October 25, 2020, from https://meridian.allenpress.com/ijmsc/article/2/3/4/33107/Swallowing-and-Speaking-Challenges-for-the-MS

Char, D., Chung, T., Mckee, A., & Pai, A. (2018, June). The human keyboard. Retrieved August 27, 2020, from https://scholarcommons.scu.edu/cgi/viewcontent.cgi?article=1043&context=idp_senior

Cohen, M. (2020, November 30). Signal processing problems, solved in MATLAB and in python [MOOC]. Udemy. Retrieved August 27, 2020, from https://www.udemy.com/course/signal-processing/.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J., & Brumberg, J. (2011, August 20). Silent speech interfaces. Retrieved August 25, 2020, from https://hal.archives-ouvertes.fr/hal-00616227

Eremenko, K., & Ponteves, H., DataScience, S. (n.d.). Machine learning A-Z: Hands-on python & R in data science [MOOC]. Udemy. Retrieved December 17, 2020, from https://www.udemy.com/course/machinelearning/

Hartman, K. (n.d.). Getting started with Myoware muscle sensor. Retrieved December 18, 2020, from https://learn.adafruit.com/getting-started-with-myoware-muscle-sensor

Kapur, A., Kapur, S., & Maes, P. (2018, March 5). AlterEgo: A personalized wearable silent speech interface. Retrieved October 25, 2020, from https://www.media.mit.edu/publications/alterego-IUI/

Kapur, A., Sarawgi, U., Wadkins, E., & Wu, M. (2019). Non-invasive silent speech recognition in multiple sclerosis with dysphonia. Retrieved August 24, 2020, from https://ml4health.github.io/2019/pdf/49_ml4h_preprint.pdf

Kareem, F., Azeem, M., & Sameh, A. (2017, October). Classification of EMG signals of lower arm (forearm\hand) motion patterns used to control robot hand movement. Retrieved December 30, 2020, from https://www.worldresearchlibrary.org/up_proc/pdf/1134-151151520140-45.pdf.

Karlik, B. (2018, July). Machine Learning Algorithms for Characterization of EMG Signals. Retrieved November 24, 2020, from, https://www.researchgate.net/publication/271298672_Machine_Learning_Algorithms_for_Characterization_of_EMG_Signals

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017, May). ImageNet classification with deep convolutional neural networks. Retrieved November 24, 2020, from https://dl.acm.org/doi/10.1145/3065386

MathWorks. (n.d.). Googlenet. Retrieved December 18, 2020, from https://www.mathworks.com/help/deeplearning/ref/googlenet.html

MathWorks. (n.d.) Classification learner app. Retrieved December 18, 2020, from https://www.mathworks.com/help/stats/classification-learner-app.html

Mohanty, S., Hughes, D., & Salathé, M. (2016, September 22). Using deep learning for image-based plant disease detection. Retrieved November 25, 2020, from https://pubmed.ncbi.nlm.nih.gov/27713752/

Muhammad, U. (2018, August 18). Pre-trained VGGNet architecture for remote-sensing image scene classification. Retrieved November 24, 2020, from https://www.researchgate.net/publication/329315021_Pre-trained_VGGNet_Architecture_for_Remote-Sensing_Image_Scene_Classification

Mwebaze, E., & Owomugisha, G. (2016, December). Machine learning for plant disease incidence and severity measurements from leaf images. Retrieved December 17, 2020, from https://www.researchgate.net/publication/313451669_Machine_Learning_for_Plant_Disease_Incidence_and_Severity_Measurements_from_Leaf_Images

Samuel, A. L. (1959, July). Some studies in machine learning using the game of checkers. Retrieved November 25, 2020, from https://ieeexplore.ieee.org/document/5392560

Schultz, T., & Wand, M. (2009, December 4). Modeling coarticulation in EMG-based continuous speech recognition. Retrieved August 24, 2020, from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.212.6271&rep=rep1&type=pdf

Scikit-learn. (n.d.). Scikit-learn machine learning in Python. Retrieved December 17, 2020, from https://scikit-learn.org/stable/

Tang, P., Wang, H., & Kwong, S. (2017, February 15). G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition. Retrieved November 25, 2020, from https://scholars.cityu.edu.hk/en/publications/gms2f(97b29fed-c9f3-4838-a817-c90a321c3fc7).html

Wood, T. (2019, May 17). F-score. Retrieved December 17, 2020, from https://deepai.org/machine-learning-glossary-and-terms/f-score.

Zia ur Rehman, M., Gilani, S. O., Waris, M. A., & Niazi, I. K. (2018, July). Stacked sparse autoencoders for EMG-based classification of hand motions: A comparative multi-day analysis between surface and intramuscular EMG. Retrieved November 24, 2020, from, https://www.researchgate.net/publication/326292263_Stacked_Sparse_Autoencoders_for_EMG-Based_Classification_of_Hand_Motions_A_Comparative_Multi_Day_Analyses_between_Surface_and_Intramuscular_EMG

The Effect of Over-the-Head Listening Devices on Noise-Induced Hearing Loss (NIHL)

Vedika Venkataramanan
duPont Manual High School
9th, Louisville, Kentucky

ABSTRACT

The purpose of this project was to further understand the impact that over-the-head listening devices have on noise-induced hearing loss (NIHL). A primary cause of NIHLis the use of over-the-head listening devices which can cause nerve damage, even though these devices aim to refine the sound quality of auditory materials. It was hypothesized that if the decibel level was increased in combination with the distance from the ear, the likelihood of nerve damage would increase due to the proximity at which the tone is amplified. One model was built to replicate the human ear and then tested using materials to simulate an ear exposed to headphones and earbuds. The model included a circuit that carried the sound from the outside of the ear model to the center where amplification was measured using a decibel meter. 10 trials were conducted at each of these levels: 80 dB, 85 dB, 90 dB, 95 dB, 100 dB, 105 dB. After all the trials were conducted, it was found that all results were significant excluding two, at which headphones and earbuds had very similar outputs. External factors that may have contributed to this include the changes in decibel level in the area the experiment was conducted. The results of this project indicated that the earbuds coherently held the ability to cause the most nerve damage and, therefore, proved the hypothesized statement for this project correct, as they were the closest to the internal ear.

Keywords: noise-induced hearing loss, earbuds, headphones

INTRODUCTION

Noise-induced hearing loss (NIHL) is one of the most prominent and widespread forms of sensorineural hearing loss. Recent studies have found that nearly 24% of U.S. adults ranging from the ages of 20 to 69 show symptoms that would indicate the possibility of noise-induced hearing loss (National Institute on Deafness and Other Communication Disorders, 2017). A primary cause of hearing loss is the use of over-the-head listening devices (headphones, earbuds, etc.) These devices are meant to make access to auditory materials clearer by amplifying the sound quality. Although, over the years, they have been found to cause nerve damage in the internal ear, resulting in hearing loss (Le, Straatman, Lea & Westerberg, 2017). The amplification of sound through these devices exposes the human ear to dangerous decibel levels, and when combined with the proximity of the device this can lead to noise-induced hearing loss.

Earbuds have been accredited with causing more damage to the internal ear, compared to headphones. Although, the lack of thorough research regarding this topic continues to put more individuals at risk. Headphones insulate external sound waves and are meant to be at greater proximity from the ear while still fulfilling the standard of convenience that they were intended for. This would lead one to deem them to be the safer option. NIHL, while primarily caused by over-the-head listening devices, does not designate a significant difference in which devices undertake the most damage. This is important to recognize because auditory nerve damage requires expensive measures to correct.

Trends in sensorineural hearing loss are common in the elderly but in recent years they have begun to spread to younger cohorts. The increased usage of over-the-head listening devices among all age groups, along with the lack of research indicating its relation to nerve

damage, overlooks the damage enacted. The question asked within this project was: "How do over-the-head listening devices contribute to noise-induced hearing loss (NIHL)?" The purpose of this project was to understand how everyday listening devices can cause damage to the auditory nerves, linking to potential sensorineural hearing loss in the form of NIHL. The hypothesis stated that if the decibel level is increased in combination with the distance of the over-the-head hearing device from the ear, the likelihood of nerve damage will be increased due to the proximity at which the tone is amplified. Within this experiment, multiple decibel levels were tested to understand the amplification level, ranging from the internal ear to the over-the-head listening devices (headphones and earbuds).

An analysis of past health data and the effects of over-the-head listening devices on NIHL (World Health Organization, 2015) was used to create an efficient classification of the effect these devices have on the auditory nerves. The expansion of research in this field incentivizes the understanding of the impact of over-the-head listening devices on everyday life, particularly when looking at nerve damage. This project has the potential to revolutionize research within the field.

METHODOLOGY

The goal of this experiment was to analyze the impact of over-the-head listening devices on the auditory nerves to understand the correlation between these devices and NIHL. In order to understand this data, a model of the human ear was built at a scale of 3 times that of the standard eardrum.

Foremost, a wooden block was used to create the base for all three variations of the model. Each variation of the model represents one of the following: the internal ear, exposure to earbuds, and exposure to headphones. For the model, three pieces of 2B pencil lead had to be sanded down on one side with 80 grit sandpaper, and then one was connected to a 9-volt battery connector, on the positive wire. In addition, another piece of lead was connected to a wire that was attached to a crocodile clip. These pieces of lead helped maintain the vibrations produced in the model to provide an accurate replication of the vibrations in the eardrum. Likewise, at the end of one side of each piece of lead, a base of 1 cm was created using hot glue to allow the lead to connect to the container.

As stated before, these pieces of lead had to be connected to a container. Within this experiment, a paper container with a diameter of 4 inches and a height of 2 inches was used. The lead pieces were positioned 3 cm apart from each other in the middle of the bottom of the paper container. A sturdy base was replicated by attaching a 3.5 by 3.5 inch piece of cardboard along the right edge of the wooden block, a clothespin was attached to the cardboard and positioned along the end. The container was then connected to the wooden base using the clothespin. In this layout, the lead pieces should be parallel to the wood base, sanded sides facing upward. A 9-volt battery was then glued on the corresponding edge of the base to complete the circuit. Once the wire was connected to the battery, both crocodile clips were connected to a 3.5mm male to male stereo cable that was attached to a small speaker. The final step in building the internal ear model was to place the third piece of lead sanded side down on top of those attached to the container. A small portion of trial and error was necessary to get the circuit working. This model was then altered to simulate exposure to both of the over-the-head listening devices (reference Appendix).

Replication of the headphone model required an additional 4 cm sponge that created the extra insulation in a headphone. For the earbud model, similar to that of the human ear, the input sound was much closer to the container because when earbuds are put into the ear they are generally approximately between 2-3 cm away from the eardrum.

After the model had been built, the testing procedure with each variation of the model required that the same tone was inputted into the model and then broadcasted through the small speaker. The tone was inputted at 80 dB, 85 dB, 90 dB, 95 dB, 100 dB, and 105 dB. The output volume was then recorded in decibels, for each trial. Per each input level tone, the experiment was run 10 times allowing for an accurate average and understanding of the resulting amplification. The data was then collected and averages were taken to understand the amplification levels of each device, which was then related to the

WHO's statistics on NIHL. T-tests were performed to understand the statistical significance of the data collected. The model in this experiment replicated the human ear to the closest degree that was doable in a home setting. This experiment was conducted within a residential setting.

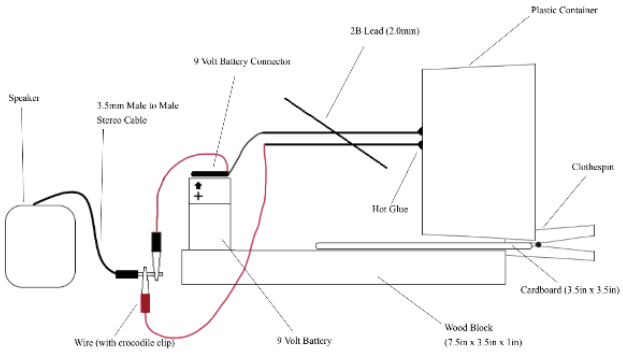


Figure 1: Model Eardrum Diagram

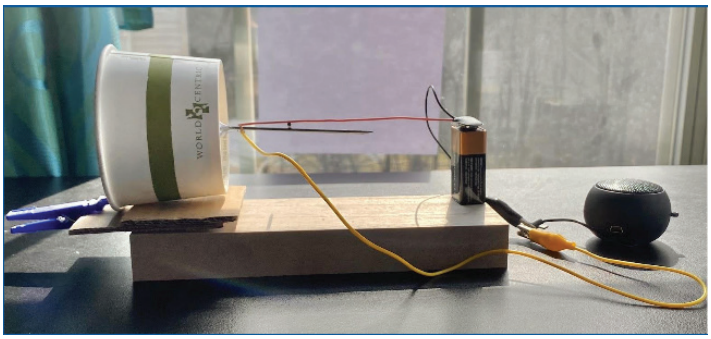


Figure 2: Model Eardrum

DATA AND RESULTS

This experiment involved three variations of a model, two of which simulated over-the-head listening devices, and one acting as the internal ear. Ten trials were performed for each variation at each decibel level input. A one-tailed uncorrelated t-test was used to compare the control to either of the additional devices as well as compare the values among those devices. The null hypothesis that was tested was that the amplified volume will be the same as that of the internal ear (control). All t-tests performed were conducted with a 0.05 level of significance and 18 degrees of freedom, which resulted from subtracting the number of trials by one.

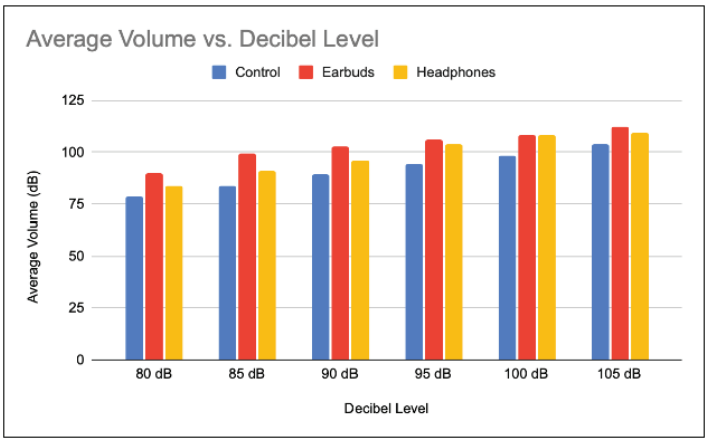


Figure 3: Average Volume vs. Decibel Level across all variations of models

Figure 3 depicts the average amplified volumes among the internal ear (control), headphones, and earbuds. The relationships between the mean values across all devices at each decibel level can be seen in the graph. The earbud variation had significantly high amplification levels consistently across decibel levels below 90 dB. At 95 dB there was only a 1.77 dB difference between the headphone and earbud outcomes, eventually leading to the headphone model having a higher amplification at 100 dB by 0.1 dB. This variation in the higher decibel levels depicts the potential effects of distortion and noise variance across mediums. This is important because earbuds are more widely available which has led to higher usage of such devices.

Table 1: *Amplification by Over-the-Head Listening Devices at 80 dB*

Over-the-head listening device	Standard Eardrum (Control)	Earbuds	Headphones	Earbuds vs. Headphones
Mean	78.71	90.09	83.65	
Variance	2.78	25.6	6.55	
Standard Deviation	1.67	5.06	2.43	
1 SD (68% Band)	77.04 - 79.38	85.03 - 95.15	81.22 - 86.08	
2 SD (95% Band)	75.37 - 82.05	79.97 - 100.21	78.79 - 88.51	
3 SD (99% Band)	73.91 - 84.51	74.91 - 105.27	76.36 - 90.94	
Number	10	10	10	
Result of t-test		t = -6.75, p < 0.00001	t = -5.12, p = 0.000036	t = 3.59, p = 0.001042
For all: df = 18, alpha = 0.05		significant	significant	significant

Table 1: T-Test for Amplification of Over-the-Head Listening Devices at 80 dB

Table 1 organizes the results and data from the t-tests conducted for the devices at 80 dB, as was done for all other decibel levels tested. The results of all three of these t-tests indicate that despite a small sample size, the data collected in this experiment has significance. The t-value for the earbuds was the highest, as negatives do not

affect the significance of a dataset. This value indicated that the data supported the research hypothesis rather than the null hypothesis. This trend was continuous through the other t-tests that were performed except for those at 95 dB and 100 dB.

The null hypothesis was supported at 95 dB and 100 dB, as the t-test between the headphones and earbuds didn't hold significance. A trend that was analyzed from both datasets collected at these levels was that the amplified mean volumes for both the headphones and earbuds were increasingly close. This was especially visible at 100 dB as it was anticipated that the null hypothesis would be accepted due to the minuscule difference of 0.1 dB between the means. This was reflected in the t-values which were both significantly lower with 1.73 at 95 dB and -0.18 at 100 dB. Therefore showing that there was a low amount of data supporting the research hypothesis. It is important to acknowledge that at 95 dB the variance across all devices was higher showing that the data was more spread out. This is evidence of the result of distortion at higher decibel levels which was accounted for only in the amplification that was measured.

Through analyzing the dataset it was visible that there were discrepancies in certain prominent trends. This did not affect the pattern across all decibel levels in which the internal ear exhibited the lowest amplification values as well as the lowest standard deviations. This further emphasized the low amounts of variability that were visible, supporting the model's efficiency in simulating the system of the internal ear. There was a significant difference for sixteen out of eighteen t-tests performed, with the main variations being presented when analyzing the data between headphones and earbuds.

CONCLUSION

The purpose of this experiment was to understand the impact that over-the-head listening devices have in contributing to (NIHL). The goal of this project was to expand this field of research by analyzing which type of over-the-head listening device amplifies sound the most when in the ear leading to nerve damage through NIHL. The hypothesis of this experiment stated that if the decibel level was increased in combination with the distance the over-the-head hearing device was from the ear, the likelihood of nerve damage would increase due to the proximity at which the tone is amplified. The primary function of these devices is to render the sound found in auditory materials. But recently, they have been found to contribute to nerve damage (Le, Straatman, Lea & Westerberg, 2017). The project was successful in analyzing the effect that over-the-head listening devices have in contributing to NIHL. The results of this experiment supported and proved the hypothesis correct. The data recorded in the amplifications had low values within the standard deviations recorded showing that there weren't any major outliers within the data. There was statistical significance for all the t-tests except two dB levels, in which the t-value was below the value appropriate for determining significance, which was 1.734. The two t-values that fell beneath this value were when testing significance between the earbuds and headphones at 95 dB with $t(18)=1.73$ and at 100 dB with $t(18)=-0.18$.

The significance of these values depicts that the data at 95 dB and 100 dB supporting the research hypothesis was highly unlikely. The mean values of the earbuds and headphones, respectively at 95 dB were 105.92 dB and 104.15 dB, showing that there is very minimal difference between the two which supports the result as insignificant. Similarly, at 100 dB, the means were 108.26 dB and 108.36 dB with only a 0.1 difference the likelihood of this test having been significant was very low. However, this was the sole instance that the headphones' mean value surpassed that of the earbuds' value. A contributing factor to this result is the potential for the difference in the surrounding decibel level during the test, as it could interfere with amplification. Although this external factor could not be eliminated in the setting in which this experiment took place, external noise does affect the amplification of sound in the ear. Therefore, in some regards, the structure of the experiment was relatively accurate since it allowed for the effect of external sound on the amplification. This enhanced the function of the model, but was not accounted for as a measured variable in the experiment.

Alternatively, it was also seen that there was a significant difference when testing earbuds

and headphones at 105 dB, the greatest level tested in this experiment to stay within potential health concerns. The t-value of 3.83 depicts that the difference between these two groups is significant and found within the results of this experiment. Despite two of the highest independent variable levels proving to have been insignificant, there is adequate information that supports earbuds are more effectively able to produce greater volumes closer to the inner ear, resulting in a higher likelihood of being exposed to high decibel levels and resulting in NIHL.

When compared to statistics provided by the World Health Organization, it has been stated that exposure to decibel levels of over 85 dB to 105 dB for over 15 minutes can cause internal damage to the ears (World Health Organization, 2015). As gleaned from this experiment, the amplified volume of over-the-head devices can grow to much higher than what is deemed as healthy. The average person listens to music using headphones or earbuds for over an hour at extended volumes (Naples & Duque, 2020). Earbuds are more commonly used than headphones and as found in this experiment are greater contributors to NIHL. Therefore, raising concerns and understanding the adequate and necessary precautions that need to be taken to avoid hearing loss is necessary.

Procedural issues must be accounted for when analyzing the results of this experiment. Using a model that was constructed after analyzing the human ear provides one of the most accurate replicas possible at this given time. That being said external noise, as well as variations in the frequency of the tone, can contribute to varied results. Alternatively, it is also necessary to acknowledge that only one type of decibel meter was used to measure all outputs. Although this provides adequate control within the experiment, decibel meters measure with slight differences.

Expanding this field of research can contribute largely to auditory sciences by analyzing the major causes of sensorineural hearing loss. Additionally, analyzing how frequency and pitch affect the reciprocity of sound through amplification can allow for a greater understanding of whether the main causes of nerve damage have to do with volume or frequency, or potentially even both. As new technologies are being developed in regards to hearing aids and devices, working towards limiting the effect of external sounds on these devices could help prevent NIHL in those who are working to sustain their hearing. By further evaluating these extensions alongside this experiment, the big picture of the overall effect of the impact of auditory devices on the internal ear can be determined.

REFERENCES

Le, T. N., Straatman, L. V., Lea, J., & Westerberg, B. (2017). Current insights in noise-induced hearing loss: a literature review of the underlying mechanism, pathophysiology, asymmetry, and management options. Journal of otolaryngology - head & neck surgery = Le Journal d'oto-rhino-laryngologie et de chirurgie cervico-faciale, 46(1), 41. https://doi.org/10.1186/s40463-017-0219-x

Naples, J., MD, & Duque, V., Au.D, CCC-A. (2020, July 22). Healthy headphone use: How loud and how long? Retrieved January 4, 2021, from https://www.health.harvard.edu/blog/healthy-headphone-use-how-loud-and-how-long-2020072220565

Noise-induced hearing loss. (2020, December 14). Retrieved January 5, 2021, from https://www.nidcd.nih.gov/health/noise-induced-hearing-loss

1.1 billion people at risk of hearing loss. (2015, August 12). Retrieved January 13, 2021, from https://www.who.int/mediacentre/news/releases/2015/ear-care/en/

U.S. adults aged 20 to 69 years show signs of noise-induced hearing loss. (2020, August 05). Retrieved September 7, 2020, from https://www.nidcd.nih.gov/news/2017/us-adults-aged-20-69-years-show-signs-noise-induced-hearing-loss

World Health Organization. (2015). Make Listening Safe [PDF file]. (2015). Geneva, Switzerland: Author. Retrieved from https://www.who.int/pbd/deafness/activities/MLS_Brochure_English_lowres_for_web.pdf