International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

A NOVEL META-MACHINE LEARNING APPROACH TO DIAGNOSE STRESS FROM INDIVIDUAL FACTORS USING A SELF-RETRIEVED DATASET AND THEN PROVIDE DIRECTED TREATMENT

Nihal Boina, Jai Agarwal, Taruna Agrawal, John Leddo, Monish Napa, Tanya Singhal, Neha Muralitharan, Poorvaja Gopinath, Kinnari Chaubal, Dillon Michlena, Ashwin Tripathy, Kalagee Mehta, Muhammad Siddiqui, SmaranPasupuleti, Riya Pasupuleti, Aneesh Sreedhara, AsvinGopinathan, HavishRallabandi, NikhitRachapudi, Nate Levkov, Alyssa Gatesman, Parth Patel,SathvikRedrouthu, JagadeepramMaddipatla, Vaneesha Gupta, Yash Sonis, Vikram Rudraraju, Aditya Sharma, Aarush Dhawan, Aastha Sharma, Saranya Ganne, Bhargav Subash, Abhiraj Tiwari, Alina Sharifi, Soureen Singh, Alec Agayan, Sauman Das, Arnav Jain, Utkarsh Goyal, Saurav Banerjee, Sadhana Mallemudi, SiddarthMallemudi, Mason Elkas, PrithamMulagura, Kavya Velaga, Nithya Jayakaran, ShriyanBachigari, VedhaBommineni, Sanaa Karkera, SharanyaChilukuri, Aaditya Panjabi, Jishnu Patel, Vihaan Cherukuri, Eeshika Singh, Inaayah Khan

MyEdMaster, LLC, 13750 Sunrise Valley Drive, Herndon, VA, United States of America

John Leddo is the director of research at MyEdmaster.

DOI: 10.46609/IJSSER.2021.v06i12.031 URL: https://doi.org/10.46609/IJSSER.2021.v06i12.031

Received: 5 Dec. 2021 / Accepted: 28 Dec. 2021 / Published: 31 Dec. 2021

ABSTRACT

One of the main goals of machine learning is to make a General Artificial Intelligence. Currently, human artificial intelligence researchers work on meticulously manipulating model parameters by hand in order to arrive at highly optimized machine learning models. In the future, a system will be needed such that a software is able to completely arrive at an optimized model to a specific topic all by itself. An increasingly aware human problem is stress, which can oftentimes lead to a variety of health issues. Artificial intelligence (AI) algorithms, specifically Random Forests, have been employed to diagnose potential mental health illnesses due to a particular personal stress. Additionally, these algorithms would be manipulated by an automated hyperparameter manipulator, using extensive machine learning to find, sort, and train, validate, and test on a dataset all by itself. Put simply, we were able to make an automated software

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

capable of making its own state-of-the-art algorithms through a meta-machine learning approach, filling the role of an AI researcher. Additionally, the software was able to achieve consistent overall testing accuracy of at least 90%, quantifying its potential use in diagnosing potential mental illnesses from survey questions identifying potential stressors. Furthermore, our software is able to go one step beyond and take steps to provide potential solutions and resources to benefit the user's condition.

Keywords: general artificial intelligence, random forests, mental illness, stress, deep learning

Introduction

Stress is commonly defined as a feeling of emotional or physical tension attributed to several factors. Common manifestations of stress include panic, irritance, annoyance, and more. Stress comes with several emotional, physical, and psychological complications. Some physical effects include the deterioration of the gut barrier, effect on bone health, cancer, and abnormal skin conditions. Some emotional effects include depression, anxiety, sleep problems, muscle aches, and muscle tension.

Stress is often attributed with staunch deadlines, constant pressure, and heavy workloads, all of which are commonplace within the tech industry. Even in the midst of awareness movements concerned with strenuous environments, the well being of developers is shrouded by a lack of job security and packages/compensation. Thus, it is no surprise that a study from Blind revealed that the stress experienced by IT workers is greater than that of frontline medical workers, with almost 60% of IT professionals admitting to feeling stressed [15]. However, stress can manifest itself differently amongst various individuals, leading to diagnosis requiring a machine learning approach.

The problem has only exacerbated in recent years, as indicated by years of data from GFI Software, which commissioned an independent survey on this topic. The survey, which started in 2012, just released its 2015 report, and found that 78% of the IT workers surveyed consider their job stressful [5]. The British Interactive Media Association (BIMA) recently revealed that tech workers are five times more likely to suffer from a mental health problem, compared to the wider population [1]. This is the question that Blind—a workplace app for tech employees—set out to answer through a user survey. The app is used by 40,000 Microsoft employees, 25,000 from Amazon, 10,000 from Google, 7,000 from Uber, 6,000 from Facebook, and thousands from other tech companies, so there is wide representation in their survey results. The one-question survey had a simple yes/no answer: "Are you currently suffering from job burnout", and 57.16%

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

answered yes [15]. A survey of 1,071 CareerCast readers found that almost 3 in 4 were experiencing higher-than-moderate stress in the workplace. Respondents were asked to rate their job stress levels on a scale of 0 to 10, with 0 denoting no stress and 10 signifying constant stress. Seventy-one percent of respondents said their jobs involved a stress score of 7 or higher. The survey found that the most common cause of stress was deadlines (30 percent) [2].

Stress can be diagnosed through blood pressure as significant levels of stress correlate to higher levels [7]. Such stress further causes humans to cope with it in unhealthy ways such as overeating, which can further lead to damage including heart attacks or strokes. Though higher blood pressure is linked to other conditions and is not necessarily always associated with stress itself, it can serve as a strong indicator [13]. Additionally, physical symptoms such as headaches can also be indicators that reveal excessive amounts of stress [14]. With the growth of the technological industry and the prominent role of IT in today's world, employees are facing the effects of stress from their work and pressures from the job itself.

Work has been done to implement machine learning to detect and analyze symptoms of stress. Machine learning can assist in detecting symptoms of stress through asking basic questions that can indicate overstressing and mental disorders.

The features for the machine learning model were carefully selected questions that directly affected stress. In order to proceed with a functional model, it is imperative to be able to quantify a given input, since the hypothesis function takes in a purely numeric vector as a parameter. The answers to these questions were therefore quantified using one hot encoding if the input was non-numerical. If not, they were simply normalized using mean normalization. Missing categorical values were filled by inserting the mode of the column vector whereas missing interval values were filled by taking the average.

Educational stress is stated to be a mental stressor caused directly or indirectly by the school. Some of these stressors include workload, pressure from parents and teachers, grade improvement, and others. High levels of educational stress can lead to physical and mental issues within the student, including but not limited to headaches, depression, fatigue, and other mood disorders. More side effects include unhealthy eating habits, having fewer self desirements, and not showing progress in academic life. Both the student and school can provide resources to improve these conditions- however, not all attempts are successful, and minor changes do not alleviate all forms of stress.

Students can resolve stress and anxiety by exercising and having strict sleeping habits. These components result in proper brain functioning, which allows the student to be more productive as

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

the brain's concentration levels arise. However, students may have the right amount of time to sleep and exercise in some cases, but they fall short on meditation and relaxation. From this abbreviation, it is critical to understand that each individual has an ingenious way of their brain's functioning.

Machine learning can detect educational stress by responding to the student's most predominant treatments and methods to reduce their stress levels. In this case, the input data would be the student's gender, age, education level, society of stay, testing positive for COVID, and aptly the responses to questions regarding binary social involvements. The entire dataset is already inputted in the machine through the Kaggle dataset implying exhaustion of student workload during the COVID-19 pandemic. The output includes the predominant evaluation of medical-based advice depending on the levels of the student's stress. In this machine, if a student inputs data of alternatives that might occur in the future and they fall under the category as a precaution for anxiety, the machine's output would be to spend a short amount of time on school work. Initially, if a student's input signs no prostration levels, then the machine's output would encourage comments to further promote the educator's well-being. If the input dictates the student's current enervation, then the machine's output would result in urgent treatments to reduce the stress according to the individual.

In this machine learning software, the evaluated estimation on the accurate stress level can be abbreviated on a scale of 0-7. Based on the input, from 0 being none, 3-4 being moderate, and seven being severe, the machine can detect the student's stress level. This aspect is efficient as to the suggestions on how to promote tranquility in the student. This allows students to further emphasize and promote self-awareness to not proliferate their mental health in future academic-related concepts. The information given from the machine is always of preliminary and manual conduct to the patient.

Mental illnesses are health conditions pressured by psychological, environmental, and biological factors. It is one of the most common illnesses, and at least one in five American adults experience it throughout their lives. Since mental illnesses are products of personal experiences, people with mental illnesses vary in age. Mental illnesses can range in severity, and some might result in medication, which may not cure the illness, but can be used to control symptoms and side effects. The consequences are wide ranging and at its worst can lead to suicide. The age-adjusted suicide rate has seen a 30% increase between 2006-2016 with a consistent increase in the past 10 years [5]. Therefore, it has become increasingly important to monitor mental illnesses and diagnose them at their earliest stages. Neuropsychiatric disorders such as schizophrenia and depression are getting more attention due to their increasing social impact [6]. With the outbreak

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

of the COVID-19 pandemic, additional attention has been placed on monitoring mental illnesses. A survey showed that 43% of the respondents in China claimed that the COVID-19 pandemic resulted in moderate to severe anxiety and depressive symptoms [7].

As the data shows, detecting mental illnesses based on symptoms is becoming increasingly important. Mental illness symptoms are generally hard to detect physically, therefore there is a need to analyze psychological data such as those from surveys accurately, without bias, which current methods fail to address.

Literature Review:

Machine learning's extensive capabilities have enabled it to be increasingly used in the medical field. By using quality data, medical machine learning technologies have impacted the field in numerous ways, such as disease diagnosis, medical imaging, surgical practices, and drug development. The uses of these data and datasets are now implemented in unsupervised learning, reinforcement learning, and supervised learning. Popular machine learning algorithms for supervised learning have been developed to improve the ability of tasks to map out inputs and outputs. For example, the Random Forest Classifier has been used often in the field of disease diagnosis. This algorithm can process essential data, make many decisions ("trees"), and combine them to create a forest. The use of machine learning with diagnosing diseases and conditions has become more and more prevalent. The National Library of Medicine Institution of Health utilized this technology to predict the stress impact on children during the COVID-19 pandemic.

We reviewed a paper called Assessment of Anxiety, Depression and Stress using Machine Learning Models. Random forest, as implied by the research paper, holds quite a bit of significance with regards to accuracy in depression and stress detection. Random forest testing consists of using J48 to create a decision tree in order to spread information gain in datasets. Tests were conducted upon the DASS42 dataset based on questionnaires between 2017-2019. Outputs consisted of five levels of severity in the categories and subcategories of stress, anxiety, and depression. The total random forest accuracy for stress here was 91.95% [16].

The application of machine learning into mental health has been around for some time and continues to become more prevalent because of the increased accessibility to data. Work has mainly been focussed on either detecting or treating mental health disorders. Datasets have varied significantly in past research. There is no standard dataset that researchers have tended to use. Several researchers prepared their own datasets with various methods including scraping text data from online interfaces such as Twitter and Facebook [2]. Many of these approaches

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

utilize Natural Language Processing algorithms to develop correlations between social media posts and stress levels [3]. Some researchers took an approach similar to ours, utilizing survey data from 3 surveys to diagnose signs of depression, and its effect in the long run. They demonstrated the feasibility of data-driven approaches when detecting depression [4]. Cosic et al. investigated mental health disorders in health care workers during the COVID-19 pandemic by developing AI algorithms. Their algorithms relied on self-reported survey data as well as more objective data from clinical records [8].

We look further into survey data for this research. Survey data can provide detailed information about a patient and provides specific information relating to the questions provided. Machine learning is also very compatible with survey data as we are essentially provided a feature with each question and map them to an output value indicating signs of mental illness. Our model further explores the correlation between education and employment and mental illnesses.

Software Components:

The software has to utilize several imports using Python to function, all listed below

1. Os
2. Tensorflow
3. Shutil
4. Open datasets
5. Selenium
6. SciPy
7. Pandas

Technology:

The software starts out by first accepting user input as to specific demographics relating to the user. Some examples of demographics may be a student, or a recently unemployed person, or a person employed in IT. Upon doing this, the software will then go to the popular machine learning website kaggle.com, and search for a related dataset pertaining to the specific related demographic through a selenium python import [7]. The software does this by entering the demographics and the subject area of mental health and specifying whether to look for datasets

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

on the Kaggle page. Following this, the software will then download the dataset all by itself using the open datasets Python import.

Once the dataset is retrieved, the software is then able to try a variety of data preprocessing methods and machine learning algorithms to finally arrive at a unique highly optimized model that it can use for the specific user-inputted disease.

A descriptive flowchart of this process is shown in Figure 1.

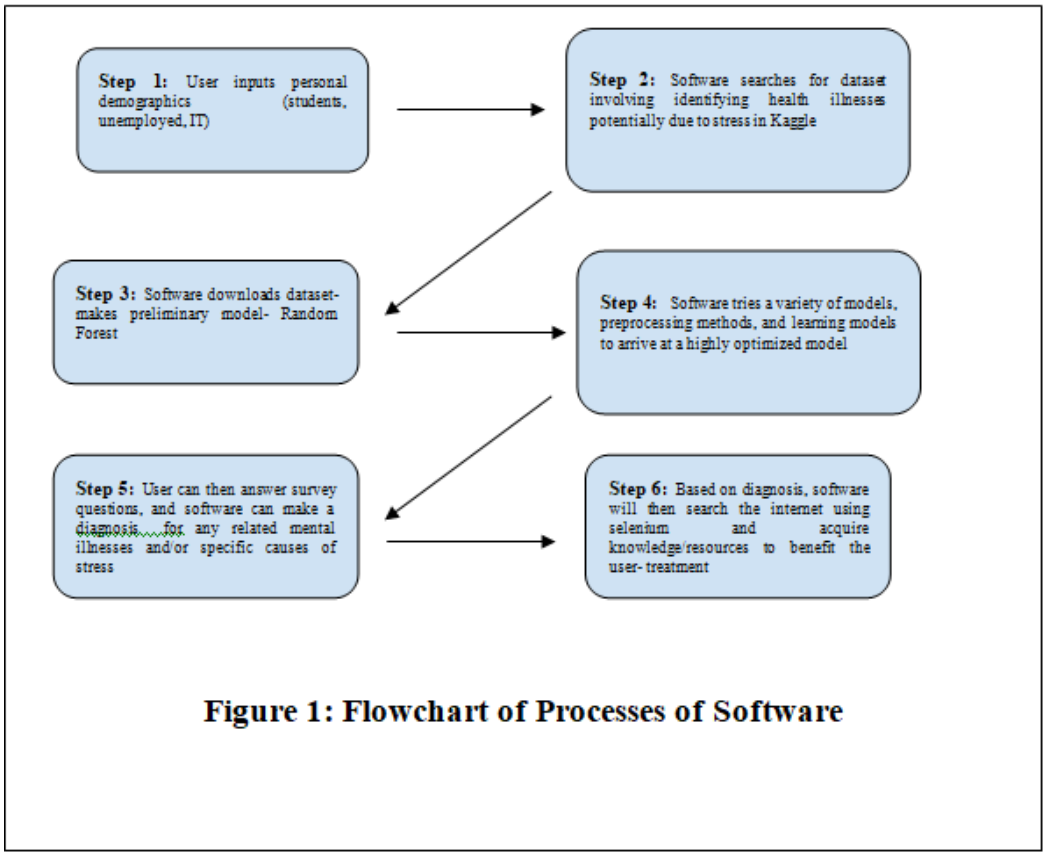

Figure 1: Flowchart of Processes of Software

Dataset:

The research leveraged three independent datasets: one for attitudes towards mental health in the tech workplace [8], one for students' stress levels in COVID-19 specifically, and another exploring the causation of high unemployment on the mentally ill.

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

The first dataset is the 2016 OSMI survey. Through 63 questions and 1434 samples pertinent to mental health conditions, the survey classified a mental health disorder if the user had one into groups. The model could be trained to be more accurate with more than 1434 data points, but those were unavailable at the time.

The second dataset is concerning the impact of an educational environment, specifically those created by COVID-19, on students' stress levels. The dataset contained 40 verified survey responses containing categorical inputs, with nominal variables such as type of environment before COVID-19, ordinal variables such as stress levels before the pandemic, and hours spent on homework and classwork. The measured outputs were stress levels due to the pandemic, measured on a scale of 0-6 from both homework and classwork. Average stress levels from the dataset increased from 3.05 to 4.38 with the addition of the pandemic, predicting a more severe level of stress. Average stress change by the state was measured as well, with states such as West Virginia and Pennsylvania experiencing the most significant changes in stress levels.

The third dataset was by NAMI (National Alliance on Mental Illness). It was a survey conducted by Michael Corley that explored the causation of high unemployment on the mentally ill. The survey, consisting of 24 questions about employment, education, and mental illness, was conducted on a group of 334 people, including 80 people with mental illnesses, which represents the current population ratio of people with mental illnesses.

Algorithm:

Since primarily surveyal data was used, a random forest classifier was primarily used on one-hot-encoded data to identify a specific mental illness from responses to a series of questions from the software's process of trying out various models. For the first dataset, it is initially preprocessed and all non-categorical data is encoded using either a one-hot encoder (OHE) or an ordinal encoder. Data is sorted into either the OHE or the ordinal pipeline based on the contents and the formatting of the OSMI data. The data is now encoded and usable to train the model. The algorithm involved a random forest classifier with depth of 20 and a random state of 42 to determine if a person was diagnosed with a stress-related mental disorder. A random forest creates multiple decision trees and takes the average of the multiple answers to come up with one conclusion. This helps achieve the highest results with the most efficiency. For the second dataset, a random forest classifier was used to take survey data and predict educational stress levels. The Kaggle dataset was split into training and testing data, splitting the inputs and outputs into "X" and "Y." There were approximately 34 training inputs and 6 for testing, which was split using the Scikit-learn library. The RandomForestClassifier was then used to fit a specific amount

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

of decision tree classifiers into the dataset's sub-samples and predict the output using machine learning, measuring the model's accuracy. An additional model was used, an XG Boost classifier, which took in the 34 inputs and six outputs and generated a prediction, displaying the model's accuracy. For the third dataset, a decision tree classifier was found to be optimal and an entropy criterion for information gain was found to be the best measure for splitting data.

Model Construction Procedure:

In order to arrive at the best possible model, the software goes through a variety of different parameters while training on 80% of the total data. There was no validation data utilized in this paper. The training data and testing data was split in a constant 4:1 ratio. After each subsequent testing accuracy, the software alternates the physical parameters in order to achieve better testing accuracy. The physical parameters that the software changes is the type of model (Decision Trees, XGBoost, AdaBoost, Multilayer Perceptron, Random Forest) as well as the activation functions, number of trees, and varying preprocessing methods. In order to then provide treatment based on a diagnosis, the software utilizes a pre-trained natural language processing (NLP) model that can extract important sentences from approved websites. This model uses an alternation of the Rapid Automatic Keyword Extraction (RAKE) algorithm with the NLTK toolkit and Selenium for web access. It does this to provide generalized information for how to better the user's condition and collect links to references for further information.

Results:

Dataset	Final Testing Accuracy
2016 OSMI Survey	95%%
Educational Stress	98%
NAMI	90%

Table 1: Final Testing Accuracies for All Models

Conclusion:

Our results show that high testing accuracies were obtained by our software, with each generated model having accuracies that were at least 90%. Our software was able to do this by itself by

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

using a random forest classifier which is optimal due to the categorical questions that were asked (which could be treated as conditional logic at each node in a subtree). This is again backed by our results, indicating that our software can accurately learn how to diagnose stress in the tech workplace, unemployment, and students through these inquiries relating to circumstances in work. Additionally, the software was able to provide helpful solutions and references to help improve the user's condition. Due to its standardized nature, his model can later be integrated with others in order to develop a more well rounded perspective on stress, which can then be worked into a larger, modular network for problem generalization. The next step for this project is to do more finetuning, to fit a variety of cases, and then implement in real life.

References

[1] Symmons, S. (2019, December 17). *Why do tech workers suffer more from mental health issues?*CPS Group UK. Retrieved November 25, 2021, from https://www.cpsgroupuk.com/blog/why-do-tech-workers-suffer-more-from-mental-health-issues.

[2] Wilkie, D. (2019, August 16). *No. 1 stressor at work: Deadlines*. SHRM. Retrieved November 25, 2021, from https://www.shrm.org/resourcesandtools/hr-topics/employee-relations/pages/workplace-stress.aspx.

[3] Centers for Disease Control and Prevention. (2014, June 6). *Stress...at work*. Centers for Disease Control and Prevention. Retrieved November 25, 2021, from https://www.cdc.gov/niosh/docs/99-101/default.html.

[4] Dishman, L. (2017, January 18). *How employee burnout became an epidemic and what it might take to fix it*. Fast Company. Retrieved November 25, 2021, from https://www.fastcompany.com/3067319/how-employee-burnout-became-an-epidemic-and-what-it-might-take-to-fix-it.

[5] *Thibodeau, P. (2015, May 8). Is it work getting more stressful, or is it the millennials? Computerworld. Retrieved November 25, 2021, from https://www.computerworld.com/article/2920309/is-it-work-getting-more-stressful-or-is-it-the-millennials.html.*

[6] Kulke, S. | B. S. K. (2020, October 29). 'Stress in America' survey reveals mental health of young adults as most at-risk. Retrieved November 25, 2021, from https://news.northwestern.edu/stories/2020/10/stress-in-america-2020-survey-reveals-mental-health-of-young-adults-as-most-at-risk/.

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

[7] Mayo Foundation for Medical Education and Research. (2021, March 18). *Stress and high blood pressure: What's the connection?* Mayo Clinic. Retrieved November 25, 2021, from https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/stress-and-high-blood-pressure/art-20044190.

[8] *Mental health statistics: Stress*. Mental Health Foundation. (2020, January 16). Retrieved November 25, 2021, from https://www.mentalhealth.org.uk/statistics/mental-health-statistics-stress.

[9] *Stress*. CAMH. (n.d.). Retrieved November 25, 2021, from https://www.camh.ca/en/health-info/mental-illness-and-addiction-index/stress#:~:text=When%20stress%20becomes%20overwhelming%20and,complaints%20such%20as%20muscle%20tension.

[10] *Technology - 57% of tech industry employees are suffering from Job Burnout*. The Helper. (n.d.). Retrieved November 25, 2021, from https://www.thehelper.net/threads/57-of-tech-industry-employees-are-suffering-from-job-burnout.166252/.

[11] U.S. Department of Health and Human Services. (n.d.). *5 things you should know about stress*. National Institute of Mental Health. Retrieved November 25, 2021, from https://www.nimh.nih.gov/health/publications/stress.

[12] Wilkie, D. (2019, August 16). *No. 1 stressor at work: Deadlines*. SHRM. Retrieved November 25, 2021, from https://www.shrm.org/resourcesandtools/hr-topics/employee-relations/pages/workplace-stress.aspx.

[13] Kulkarni S, O'Farrell I, Erasi M, Kochar MS. Stress and hypertension. WMJ. 1998 Dec;97(11):34-8. PMID: 9894438.

[14] Yaribeygi, H., Panahi, Y., Sahraei, H., Johnston, T. P., &Sahebkar, A. (2017). The impact of stress on body function: A review. EXCLI journal, 16, 1057–1072. https://doi.org/10.17179/excli2017-480

[15] Bradford, L. (2021, July 12). *Why we need to talk about burnout in the Tech Industry*. Forbes. Retrieved November 25, 2021, from https://www.forbes.com/sites/laurencebradford/2018/06/19/why-we-need-to-talk-about-burnout-in-the-tech-

ISSN: 2455-8834

Volume:06, Issue:12 "December 2021"

industry/#:~:text=How%20Common%20Is%20Employee%20Burnout,answer%20through%20a%20user%20survey.

[16] Kumar, P., Garg, S., & Garg, A. (2020). Assessment of anxiety, depression and stress using machine learning models. Procedia Computer Science, 171, 1989–1998. https://doi.org/10.1016/j.procs.2020.04.213

