Hamido Fujita
Ali Selamat
Jerry Chun-Wei Lin
Moonis Ali (Eds.)

LNAI 12798

Advances and Trends in Artificial Intelligence

Artificial Intelligence Practices

34th International Conference
on Industrial, Engineering and Other Applications
of Applied Intelligent Systems, IEA/AIE 2021
Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I

Springer

Lecture Notes in Artificial Intelligence 12798

Subseries of Lecture Notes in Computer Science

More information about this subseries at http://www.springer.com/series/1244

Hamido Fujita · Ali Selamat ·
Jerry Chun-Wei Lin · Moonis Ali (Eds.)

Advances and Trends in Artificial Intelligence

Artificial Intelligence Practices

34th International Conference
on Industrial, Engineering and Other Applications
of Applied Intelligent Systems, IEA/AIE 2021
Kuala Lumpur, Malaysia, July 26–29, 2021
Proceedings, Part I

Springer

Editors
Hamido Fujita
i-SOMET Incorporate Association
Morioka, Japan

Ali Selamat
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia

Jerry Chun-Wei Lin
Western Norway University
of Applied Sciences
Bergen, Norway

Moonis Ali
Texas State University San Marcos
San Marcos, TX, USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-030-79456-9 ISBN 978-3-030-79457-6 (eBook)
https://doi.org/10.1007/978-3-030-79457-6

LNCS Sublibrary: SL7 – Artificial Intelligence

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Artificial Intelligence innovations in recent decades have entered a sophisticated stage in providing intelligent interaction between humans and machines, solving problems, and providing advice in many different infrastructures. Machines in different disciplines have become ubiquitous in all aspects of life, including education, governance, science, healthcare, warfare, and industry. Computing machinery has become smaller and faster, and the costs of data storage and communication have greatly decreased. Consequently, big data of vast dimensionality is being intelligently collected and stored in smart databases for use in decision making and prediction for applications such as security and health care, amongst others. Moreover, novel and improved computing architectures have been designed for efficient large-scale data processing, such as big data frameworks, FPGAs and GPUs. Thanks to these advancements and recent breakthroughs in artificial intelligence, researchers and practitioners have developed more complex and effective artificial intelligence-based systems. This has led to a greater interest in artificial intelligence to solve complex real-world problems, and the proposal of many innovative applications.

This volume contains the proceedings of the 34th International Conference on Industrial, Engineering and other Applications of Applied Intelligent Systems (IEA/AIE 2021), which was held online during July 26–29, 2021, in Kuala Lumpur, Malaysia. The IEA/AIE conference is an annual event that emphasizes applications of applied intelligent systems to solve real-life problems in all areas including engineering, science, industry, automation and robotics, business and finance, medicine and biomedicine, bioinformatics, cyberspace, and human-machine interactions. This year, 145 submissions were received. Each paper was evaluated by three to four reviewers from an International Program Committee consisting of 196 members from 37 countries. Based on the evaluation, 87 papers were selected as full papers and 19 as short papers, which are presented in two volumes. We are grateful to all the reviewers for the time spent writing detailed and constructive comments for the authors, and also the authors for the proposal of so many high-quality papers.

The program of IEA/AIE 2021 included eight special sessions:

- Special Session on Data Stream Mining: Algorithms and Applications (DSMAA2021)
- Special Session on Intelligent Knowledge Engineering in Decision Making Systems (IKEDS2021)
- Special Session on Knowledge Graphs in Digitalization Era (KGDE2021)
- Special Session on Spatiotemporal Big Data Analytics (SBDA2021)
- Special Session on Big Data and Intelligence Fusion Analytics (BDIFA2021)
- Special Session on AI in Healthcare (AIH2021)
- Special Session on Intelligent Systems and e-Applications (iSeA2021)
- Special Session on Collective Intelligence in Social Media (CISM2021).

Moreover, two keynote talks were given by Professor Francisco Herrera, from the University of Granada, Spain, and Director of the Andalusian Research Institute "Data Science and Computational Intelligence", and Professor Vincent S. Tseng from the Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan.

We would like to thank everyone who has contributed to the success of this year's edition of IEA/AIE, that is the authors, Program Committee members, reviewers, keynote speakers, organizers and participants.

May 2021

Hamido Fujita
Ali Selamat
Jerry Chun-Wei Lin
Moonis Ali

Organization

General Chairs

Hamido Fujita, Japan
Moonis Ali, USA

Organizing Chairs

Ali Selamat, Malaysia
Jun Sasaki, Japan

Program Chairs

Ali Selamat, Malaysia
Jerry Chun-Wei Lin, Norway

Special Session Chairs

Philippe Fournier-Viger, China
Nor Azura Mohd Ghani, Malaysia

Publicity Chairs

Mohd Hazli Mohamed Zabil, Malaysia
Lim Kok Cheng, Malaysia

Program Committee

Abidalrahman Moh'D, USA
Adel Bouhoula, Tunisia
Adrianna Kozierkiewicz, Poland
Ahmed Tawfik, Egypt
Alban Grastien, Australia
Alexander Ferrein, Germany
Artur Andrzejak, Germany
Ayahiko Niimi, Japan
Barbara Pes, Italy
Bay Vo, Vietnam
Dariusz Krol, Poland
Dinh Tuyen Hoang, Korea
Du Nguyen

Engelbert Mephu Nguifo, France
Eugene Santos Jr., USA
Farid Nouioua, France
Farshad Badie, Denmark
Fevzi Belli, Germany
Flavio Soares Correa da Silva, Brazil
Franz Wotawa, Austria
Giorgos Dounias, Greece
Hadjali Allel, France
Hamido Fujita, Japan
He Jiang, China
Ingo Pill, Austria
Jerry Chun-Wei Lin, Norway
Joao Mendes-Moreira, Portugal
João Paulo Carvalho, Portugal
Jose Maria-Luna, Spain
Krishna Reddy P., India
Ladjel Bellatreche, France
Leszek Borzemski, Poland
Maciej Grzenda, Poland
Mark Levin, USA
Mercedes Merayo, Spain
Nazha Selmaoui-Folcher, Germany
Ngoc-Thanh Nguyen, Poland
Philippe Fournier-Viger, China
Philippe Leray, France
Rui Abreu, Portugal
Sabrina Senatore, Italy
Said Jabbour, France
Shyi-Ming Chen, Taiwan
Sonali Agarwal, India
Takayuki Ito, Japan
Tim Hendtlass, Australia
Trong Hieu Tran, Vietnam
Tzung-Pei Hong, Taiwan
Uday Rage, Japan
Unil Yun, Korea
Van Cuong Tran, Vietnam
Wen-Juan Hou, Taiwan
Wolfgang Mayer, Australia
Xiangdong An, USA
Xinzheng Niu, China
Yun Sing Koh, Australia
Yutaka Watanobe, Japan

Contents – Part I

Knowledge Discovery and Pattern Mining

Fast Mining of Top-k Frequent Balanced Association Rules 3
Xiangyu Liu, Xinzheng Niu, Jieliang Kuang, Shenghan Yang, and Pengpeng Liu

Towards Increasing Open Data Adoption Through Stream Data Integration and Imputation 15
Robert Kunicki and Maciej Grzenda

Towards Efficient Discovery of Periodic-Frequent Patterns in Columnar Temporal Databases 28
Ravikumar Penugonda, Likhitha Palla, Uday Kiran Rage, Yutaka Watanobe, and Koji Zettsu

Data-Driven Simulation of Ride-Hailing Services Using Imitation and Reinforcement Learning 41
Haritha Jayasinghe, Tarindu Jayatilaka, Ravin Gunawardena, and Uthayasanker Thayasivam

Discovering Spatial High Utility Itemsets in High-Dimensional Spatiotemporal Databases 53
Sai Chithra Bommisetty, Ravikumar Penugonda, Uday Kiran Rage, Minh Son Dao, and Koji Zettsu

A Single-Stage Tree-Structure-Based Approach to Determine Fuzzy Average-Utility Itemsets 66
Tzung-Pei Hong, Meng-Ping Ku, Hsiu-Wei Chiu, Wei-Ming Huang, Shu-Min Li, and Jerry Chun-Wei Lin

Mining Episode Rules from Event Sequences Under Non-overlapping Frequency 73
Oualid Ouarem, Farid Nouioua, and Philippe Fournier-Viger

Distributed Mining of High Utility Time Interval Sequential Patterns with Multiple Minimum Utility Thresholds 86
Sumalatha Saleti, Jaya Lakshmi Tangirala, and Ragunathan Thirumalaisamy

Computational Ontology and BIM Technology in Data-Driven Indoor Route Planning 235
Barbara Strug and Grażyna Ślusarczyk

Collaborative Maintenance of EDOAL Alignments in VocBench 243
Armando Stellato, Manuel Fiorelli, Tiziano Lorenzetti, and Andrea Turbati

DIKG2: A Semantic Data Integration Approach for Knowledge Graphs Generation from Web Forms 255
Rahma Dandan and Sylvie Despres

Ontology-Based Resume Searching System for Job Applicants in Information Technology 261
Tung T. Phan, Vinh Q. Pham, Hien D. Nguyen, Anh T. Huynh, Dung A. Tran, and Vuong T. Pham

An Approach to Expressing Metamodels' Semantics in a Concept System . . . 274
Marcin Jodłowiec and Marek Krótkiewicz

Medical and Health-Related Applications

Birth-Death MCMC Approach for Multivariate Beta Mixture Models in Medical Applications 285
Mahsa Amirkhani, Narges Manouchehri, and Nizar Bouguila

Intelligent Asthma Self-management System for Personalised Weather-Based Healthcare Using Machine Learning 297
Radiah Haque, Sin-Ban Ho, Ian Chai, Chin-Wei Teoh, Adina Abdullah, Chuie-Hong Tan, and Khairi Shazwan Dollmat

Deep Forecasting of COVID-19: Canadian Case Study 309
Fadoua Khennou and Moulay A. Akhloufi

COVID-19 Genome Analysis Using Alignment-Free Methods 316
M. Saqib Nawaz, Philippe Fournier-Viger, Xinzheng Niu, Youxi Wu, and Jerry Chun-Wei Lin

Deep Efficient Neural Networks for Explainable COVID-19 Detection on CXR Images 329
Mohamed Chetoui and Moulay A. Akhloufi

Predicting Psychological Distress from Ecological Factors: A Machine Learning Approach 341
Ben Sutter, Raymond Chiong, Gregorius Satia Budhi, and Sandeep Dhakal

Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)

COUNTRY

Germany

Universities and research institutions in Germany

Media Ranking in Germany

SUBJECT AREA AND CATEGORY

Computer Science
└ Computer Science (miscellaneous)

Mathematics
└ Theoretical Computer Science

PUBLISHER

Springer Science and Business Media Deutschland GmbH

SJR 2024

0.352 Q2

H-INDEX

499

PUBLICATION TYPE

Book Series

ISSN

03029743, 16113349

COVERAGE

1973-2025

INFORMATION

Homepage

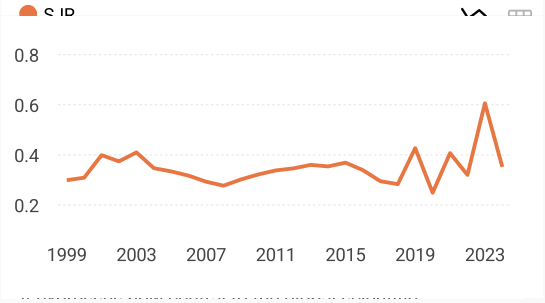
How to publish in this journal

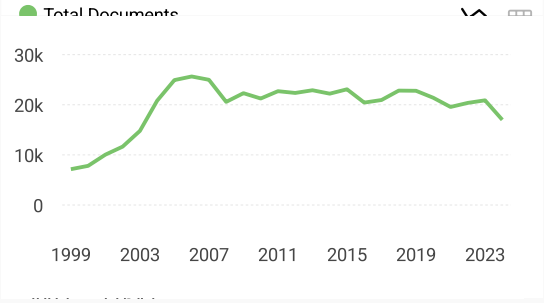
lncs@springer.com

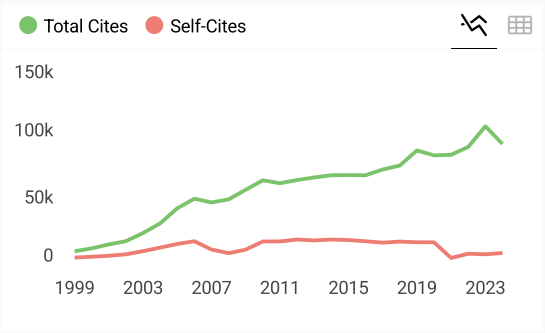
SCOPE

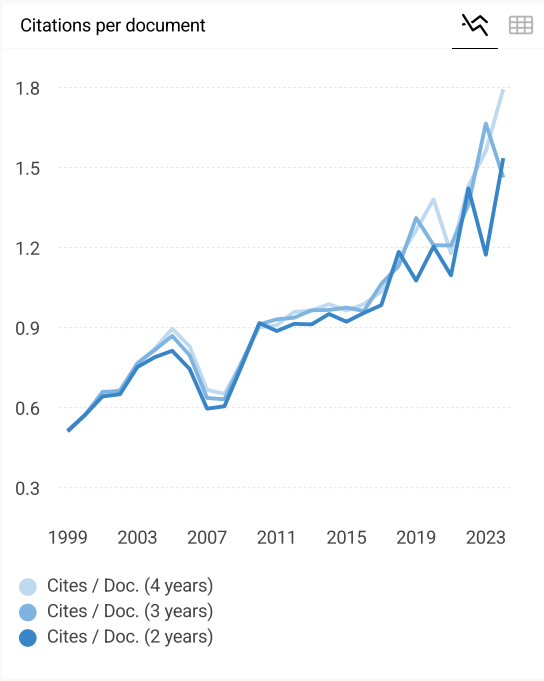
This distinguished conference proceedings series publishes the latest research developments in all areas of computer science.

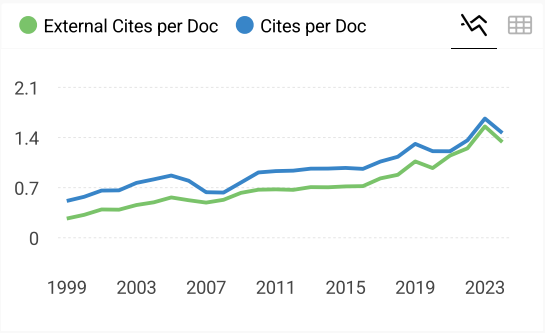
Quartiles

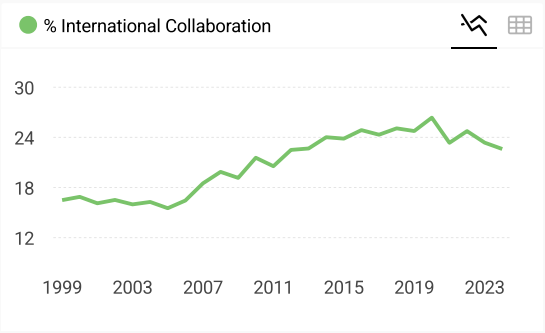

SJR
0.8
0.6
0.4
0.2
1999 2003 2007 2011 2015 2019 2023

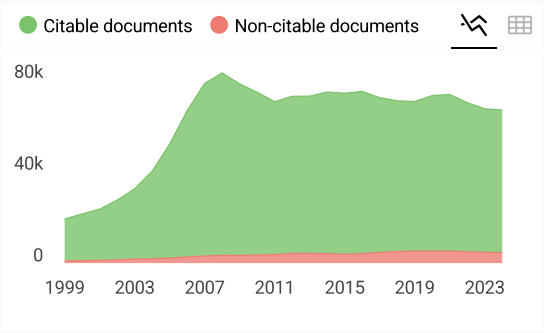

Total Documents
30k
20k
10k
0
1999 2003 2007 2011 2015 2019 2023

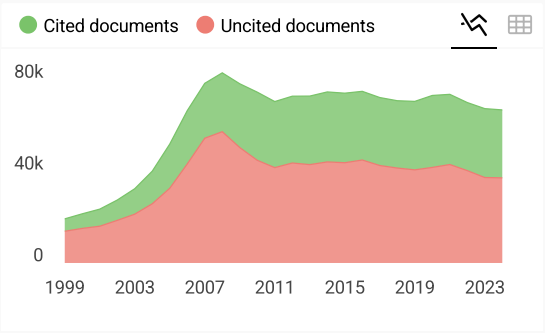

Total Cites
Self-Cites
150k
100k
50k
0
1999 2003 2007 2011 2015 2019 2023


Citations per document
1.8
1.5
1.2
0.9
0.6
0.3
1999 2003 2007 2011 2015 2019 2023
Cites / Doc. (4 years)
Cites / Doc. (3 years)
Cites / Doc. (2 years)

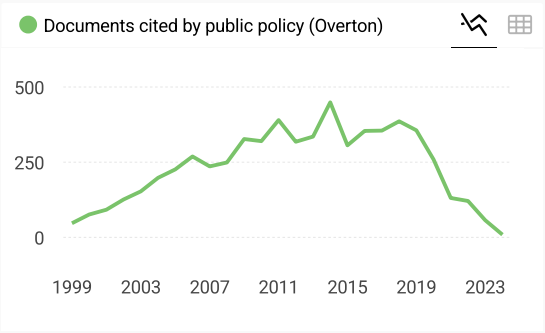

External Cites per Doc
Cites per Doc
2.1
1.4
0.7
0
1999 2003 2007 2011 2015 2019 2023

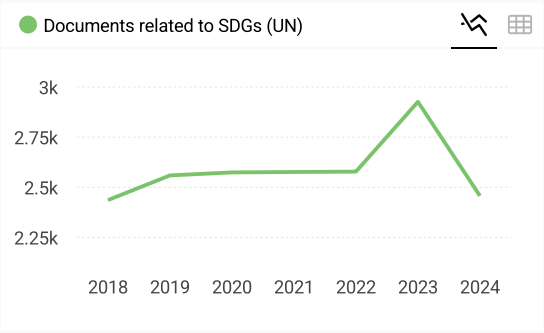

% International Collaboration
30
24
18
12
1999 2003 2007 2011 2015 2019 2023

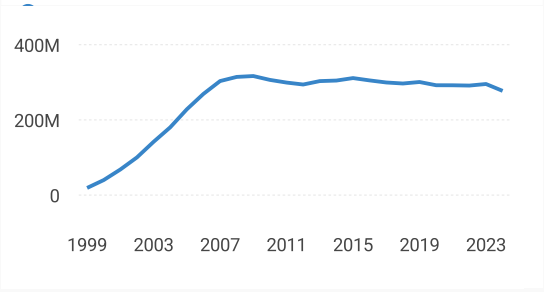

Citable documents
Non-citable documents
80k
40k
0
1999 2003 2007 2011 2015 2019 2023

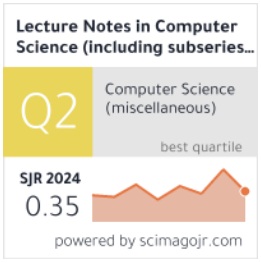

Cited documents
Uncited documents
80k
40k
0
1999 2003 2007 2011 2015 2019 2023


% Female Authors
32
24
16
8
1999 2003 2007 2011 2015 2019 2023


Documents cited by public policy (Overton)
500
250
0
1999 2003 2007 2011 2015 2019 2023


Documents related to SDGs (UN)
3k
2.75k
2.5k
2.25k
2018 2019 2020 2021 2022 2023 2024





← Show this widget in your own website

Just copy the code below and paste within your html code:

```
<a href="https://www.scimag
```

SCImago Graphica

Explore, visually communicate and make sense of data with our **new data visualization tool**.

Metrics based on Scopus® data as of March 2025

Jason Tan 5 months ago

May I know why when I searched on this site,
https://www.scopus.com/sources.uri

LNCS is ranked at 48% which is not Q2?

reply

SCImago Team

Melanie Ortiz 5 months ago

Dear Jason,
Thank you for contacting us.
As you probably already know, our data come from Scopus, they annually send us an update of the data. This update is sent to us around April / May every year.
The calculation of the indicators is performed with the copy of the Scopus database provided to us annually. Regarding your inquiry about the Quartile distribution process at SCImago, the journals are ranked and distributed in 4 equal groups based on their SJR value, unlike Scopus, who ranks the publications by percentiles based on the journal's CiteScore.
The Quartile methodology, like others that are used to group results such as percentiles, can be applied to any indicator. Currently, Scopus offers information on the journals ranking and the percentile they occupy according to the CiteScore indicator (https://service.elsevier.com/app/answers/detail/a_id/14880/supporthub/scopus/), which is perceived as an impact indicator, but that is different from the SJR, as the latter is also a normalized impact indicator (https://www.scimagojr.com/files/SJR2.pdf).
Both Scopus and SCImago Journal and Country Rank offer information on the SJR indicator for every journal, although the position of each of the publications and the quartile in which it is located according to the SJR can be consulted at https://www.scimagojr.com.
According to the above, the difference in the information consulted on the Scopus journal's profile and in Scimagojr.com lies in the fact that they represent the position of

Source details

Lecture Notes in Computer Science

Open Access

Years currently covered by Scopus: from 1973 to 2026

Publisher: Springer Nature

ISSN: 0302-9743 E-ISSN: 1611-3349

Subject area: Mathematics: Theoretical Computer Science | Computer Science: General Computer Science

Source type: Book Series

CiteScore 2024
2.4

SJR 2024
0.352

SNIP 2024
0.555

View all documents > | Set document alert | Save to source list

CiteScore | CiteScore rank & trend | Scopus content coverage

CiteScore 2024

$$2.4 = \frac{175{,}620 \text{ Citations } 2021 - 2024}{72{,}867 \text{ Documents } 2021 - 2024}$$

Calculated on 05 May, 2025

CiteScoreTracker 2025

$$2.5 = \frac{187{,}545 \text{ Citations to date}}{75{,}382 \text{ Documents to date}}$$

Last updated on 05 October, 2025 • Updated monthly

CiteScore rank 2024

Category	Rank	Percentile
Mathematics Theoretical Computer Science	#77/136	43rd
Computer Science General Computer Science	#139/239	42nd

View CiteScore methodology > | CiteScore FAQ > | Add CiteScore to your site

Predicting Psychological Distress from Ecological Factors: A Machine Learning Approach

Ben Sutter[1], Raymond Chiong[1(✉)], Gregorius Satia Budhi[1,2], and Sandeep Dhakal[1]

[1] School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW 2308, Australia
Chiong@newcastle.edu.au

[2] Informatics Department, Petra Christian University, Surabaya 60236, Indonesia

Abstract. Over 300 million people worldwide were suffering from depression in 2017. Australia alone invests more than $9.1 billion each year on mental health related services. Traditional intervention methods require patients to first present with symptoms before diagnosis, leading to a reactive approach. A more proactive approach to this problem is highly desirable, and despite ongoing work using approaches such as machine learning, further work is required. This paper aims to provide a foundation by building a machine learning model across multiple techniques to predict psychological distress from ecological factors alone. Eight different classification techniques were implemented on a sample dataset, with the best results achieved through Logistic Regression, providing an accuracy of 0.811. The preliminary results suggest that, with future improvements on implementation and analysis, an accurate and reliable model is possible. This study, with the proposed base model, can potentially lead to the development of a proactive solution to the global mental health crisis.

AQ1

1 Introduction

The World Health Organisation (WHO), in 2017, reported that more than 300 million people worldwide – ≈4.4% of the global population – were suffering from depression [34,37]. Similarly, according to the Australian Institute of Health and Welfare, $9.1 billion was spent on mental health-related services in 2016-17, and 2.5 million people (≈10% of the Australian population) received Medicare-subsidised mental health-specific services in 2017–18 [1]. These statistics highlight the severity and the widespread nature of mental health issues, and with the growing awareness of the problem, there has been a significant increase in research and funding for the detection and prediction of mental health issues.

Leightley et al. [21], while focusing on the identification of post-traumatic stress disorder (PTSD) in a United Kingdom military cohort, also assessed the impact of mental health on the day-to-day duties of serving and ex-serving soldiers, specifically on their retention and productivity. Similarly, Walsh et al. [36]

H. Fujita et al. (Eds.): IEA/AIE 2021, LNAI 12798, pp. 1–12, 2021.
https://doi.org/10.1007/978-3-030-79457-6_30

outlined the significance of psychological distress in adolescents, with suicide being the second leading cause of death in adolescents. For each suicide in the United States, there are 100–200 non-fatal attempts [36]. Mental health and psychological distress is a global issue, which costs our health care systems billions of dollars each year, and it is clearly non-discriminatory.

The application of machine learning (ML) approaches towards mental health and psychological distress problems is an ongoing research endeavour. Several studies have successfully built prediction models for psychological distress using ML techniques such as the Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbour (KNN), Decision Tree (DT) and Random Forest (RF) [13,23,28,29,34,36]. Despite the ever increasing of research on improving mental health and psychological distress diagnosis with ML, the reoccurring theme, however, is the use of historical records or user reported surveys to train the ML models. Although different ML classification techniques can be used to accurately predict psychological distress, a vast majority of them rely on people self-presenting for assessment, or self-identifying their condition before the key features are available for analysis [20,34]. There is, thus, a void with regards to generalised prediction from ecological factors alone [16]. We propose that the use of ecological factors would provide a proactive approach to generalised prediction. The few studies that have been conducted on ecological factors (e.g., see [25]) are based only on formulated questionnaire responses rather than scrutinised psychological assessment and screening tools. Therefore, this study aims to bridge this gap in the literature and supplement existing modelling research by providing a strategy to predict psychological distress based on ecological factors.

More specifically, the primary objective here is to bridge the gap between real-time ecological factors and existing psychological distress research. For this to be successful, the ML model should accurately and reliably categorise a specific person's psychological distress based solely on their ecological factors. All measurements, such as the accuracy, precision, recall, F-measure (F1), and area under the curve (AUC), should be comparable to or outperform similar ML techniques in the referenced literature. It should be noted that, in the context of this study, recall must give high scores. Failure to predict positive cases of psychological distress would be a major disadvantage. If successful, the developed model could be supplemented by other ML classification techniques, or used independently in real-time software to predict and report psychological distress, providing a proactive rather than reactive approach to mental health. Ultimately, a proactive approach could then be used to offer alternate content, or even to alert a third party to provide more intense intervention methods, before the person reaches the state of potential self-harm or suicide.

Trotzek et al. [34] conducted an exhaustive literature review to identify the ecological risk factors for PTSD. They used this information to develop a questionnaire, which was then used to generate the dataset for their study. Once ecological risk factors and psychometric properties of a questionnaire are established, further research is generally required to validate the questionnaire and

verify its performance against existing questionnaire screening tools. Considering the limited time frame of our study, rather than devising a new screening tool, an ML model was developed to predict screening results based on existing and real-time ecological risk factors.

The K10 and K6 are two screening scales commonly utilised for assessing psychological distress [20], and both contain the psychometric properties required to quickly and efficiently categorise a person's psychological distress. Although other screening tools exist in the literature, such as the General Health Questionnaire (GHQ-12), research has shown that the K10 and K6 surveys perform better and are more informative in ruling in or out target disorders [9,15]. Considering that the K6 survey is a reduced version of the K10 survey (using six of the ten questions), this study uses only the K10 screening scale. The aim here is to propose an ML-based model to efficiently predict a K10 score, or psychological distress classification based on ecological factors. This base model can potentially be extended to incorporate other ML aspects such as facial recognition [28] and text analysis [34] to further enhance its efficiency and effectiveness as a real-time, proactive prediction tool. It can also be easily included as part of a real-time tool or mobile application for predicting psychological distress.

The rest of this paper is organised as follows. In Sect. 2, we first discuss the relevant literature and how existing research has contributed to the psychological prediction space. The research methodology used in this study is then described in Sect. 3, and the results of the study are presented in Sect. 4. Finally, we draw our conclusion in Sect. 5, along with proposed future work.

2 Related Work

Mor et al. [25] conducted a study in 2018 to evaluate an ML approach for identifying individuals at risk for PTSD using ecological risk factors. Initially, they generated a list of ecological risk factors, which resulted in a 37-question survey. The questionnaire was distributed to 1,290 residents of southern Israel who had been exposed to terror attacks. An ML model was then trained – using 10-fold cross-validation – on the provided ecological risk factors with a value whether or not the study participants had previously reported a PTSD diagnosis. Their model yielded the best results of AUC = 0.91 and F1 score = 0.83. This study was one of the few that included ecological factors. Even though the study does use ecological factors for assessment, it must be noted that these factors were assessed in the context of the study itself, and then used to assess a population of the same specific demographic. Although good results were achieved, the model could have been validated in a general manner by utilising commonly scrutinised psychological assessment tools and applying over a more generalised demographic.

A similar study in 2019 screened a total of 470 seafarers for anxiety and depression using ML [33]. This study also used a range of ecological factors such as age, educational qualifications, marital status and income as feature inputs to target a known Hamilton Anxiety and Depression rating. Results of 5 classification techniques produced high accuracy (>0.75) and AUC scores (>0.8, except

for the SVM with 0.759) [33]. With accurate predictions, this study successfully predicted anxiety and/or depression from ecological factors. Results could have been further validated by including a control set of people from the general population, outside of the same occupation and demographic to that of the seafarers.

Kessler et al. [20] tested ML algorithms to predict the persistence and severity of major depressive disorder. This study consisted of an initial survey of 5,877 participants, and then a re-survey of 5,001 of those participants 10–12 years later. The study used ensemble regression trees and 10-fold cross-validated penalised regression to generate a model, which was then compared against the self-reported results 10–12 years after the baseline. The study resulted in 34.6–38.1% of respondents with high persistence and 40.8–55.8% with severity indicators being in the top 20% of the baseline ML predicted distribution. Interestingly, the ML model also showed that 20% of respondents with lowest predicted risk account for only 0.9% of all hospitalisations, resulting in a prediction model useful for both high risk prediction and ruling out low risks. This study successfully outlined the benefits of using ML algorithms in psychological prediction. Re-assessing after 10–12 years allowed the ML model to be validated against real-world data, instead of just data subsets. The disadvantage of this approach, however, is the requirement of self-assessment for reporting; because, with self-assessment, there is no way of validating whether a respondent is reporting based on prior diagnosis, reporting false diagnosis or failing to report positive diagnosis.

In 2020, Priya et al. [29] applied ML algorithms to predict anxiety, depression and stress in modern life. They focused on these mental health factors by collecting results of the Depression, Anxiety and Stress Scale questionnaire (DASS 21) of 348 participants of varying age, gender and demographic. The questionnaire results were then classified using the DT, RF, NB, SVM and KNN models, with results ultimately measured by F1. The dataset was divided 70:30 into training and testing subsets. NB classification resulted in the best overall accuracy, with anxiety, depression and stress ranging between 0.73-0.85. RF classification, however, produced the best F1 scores (0.47–0.76). Similar to the work of Kessler et al. [20], all classification models also produced good results for negative cases. However, this study focused specifically on the self-reported DASS 21 questionnaire – although accurate models can be trained, comparing results against generalised ecologically inspired models is difficult in practice.

Trotzek et al. [34] addressed the early detection of depression using ML models based on messages published on the social platform, Reddit. The study compiled a range of 10 to 2,000 messages collected from a total of 135 depressed users and a random control group of 752 users [34]. The 135 depressed users were identified as depressed by posting language such as “I was diagnosed with depression”. As with other studies based on self-reporting or self-diagnosis, such identification puts the validity of these messages into question. Without any context of the message, or some form of sentiment analysis (e.g., see [7]), it is possible that people in the depressed category were posting negligently, or

people in the control group who may actually be depressed simply did not use depressive language in their comments.

A 2018 study by Walsh et al. [36] aimed to use ML to predict suicide attempts in adolescents. This retrospective study used data from 974 adolescents with non-fatal suicide attempts, 496 adolescents with other self-injury, 7,059 adolescents with depressive symptoms, and 25,081 adolescent general hospital controls [36]. Using a range of ML classification techniques, some accurate predictive models were found. Although ecological factors were not prioritised, this study still outlined the significance of medical history in prediction analysis, and suggested that a generalised model should utilise a holistic approach.

Studies have also been conducted in psychological distress prediction by analysing MRI images [23], relating whole-brain activity patterns to facial expressions [28], and further analysis of text-based comments on social platforms [13]. All these contribute to the research space; however, they fail to fill the void in research around proactive prediction without relying on historical data. Understandably, supervised ML requires historical data to train models, so it will always play a role in this field of research. The gap in the literature, and one that this study aims to address, is to use these known classification techniques to create a generalised prediction model based on ecological factors entirely.

Numerous studies on psychological distress prediction have also been done outside the ML domain. For example, Brooks et al. [3] conducted a study on self-reported psychological distress following a concussion incident among children and adolescents. Participants were assessed 4 and 12 weeks post-concussion using multiple psychological categorisation scales, and logistic regressions were used for prediction. Loula and Monteiro adopted a game theory-based model for predicting depression due to frustration in competitive environments [22]. This study introduced a game, relating investment in formal education to professional success, and proposed that an individual becomes depressed when the difference in their earnings and those of their neighbours in the game is above a threshold. Despite the research outside the ML domain, we have chosen ML in this study because of the motivating examples and existing work in the ML field, as well as the potential for future work in using a trained model in real time applications.

3 Methods

This work was initially broken into three phases: (1) Dataset and Targets, (2) Model Creation, and (3) Classification Analysis.

3.1 Dataset and Targets

In order to test the performance of the model proposed in this study, we used a public dataset by Every-Palmer et al. [12], which consists of 2 numeric and 15 categorical features, as shown in Table 1. This dataset was chosen because it contains quantifiable ecological factors as well as a K10 score, and is therefore highly suited for the purpose of our study.

Table 1. Normalised dataset dictionary

Type	Variable	Variable detail
Categorical	gender_fct3	Gender
Categorical	eth_fct4	Prioritised Ethnicity
Categorical	r3.2_livealone_fct2	Participant lives alone
Categorical	r3.4_fct2	Happiness with bubble
Categorical	r3.6_fct4	Easy to maintain contact with friends and family
Categorical	r3.9	Family relationships
Categorical	r3.10_fct2	Poor relationships with other occupants (of house)
Categorical	r3.11_fct3	Loneliness
Categorical	r4.1_fct6	Employment status
Categorical	r4.5_fct3	Type of work (essential, non-essential worker)
Categorical	r5.2_fct2	Self-rated health
Categorical	r5.3_incpreg_fct3	Health vulnerabilities for COVID (including pregnancy)
Categorical	r8.17_fct2	Prior mental health diagnosis
Categorical	r11.2_anyfamilyharm_fct2	Any reported family harm in lockdown
Categorical	r11.3_fct3	Witnessed any reported family harm in lockdown
Numeric	age_num	Age (in years)
Numeric	r6.4_num	Alcohol intake (pre lockdown)
Target	r8.6_k10_num	K-10 score (numeric)

A number of metrics from the dataset were dropped from our study, such as internal identifiers and duplicate groupings – a numeric age metric was used instead of the categorical age range grouping. Given that we used only the K10 score in this study, the remaining psychological distress scales were also dropped, including the WHO-5 and GAD-7 scores. Specific COVID-19 metrics, such as infection and test results, were also dropped, since we wanted to generalise our study outside of the COVID-19 context. Even though the proposed model would have trained successfully with the data, the aim of this work is to create a generalised psychological prediction model. Scaling was used on the age and alcohol consumption numeric metrics. One-hot encoding was used to normalise the remaining categorical metrics, which mostly consist of 3 or 4 pre-determined string formatted answers. Before normalisation, any missing data in a categorical column was replaced with the string "no value" to prevent exception. Following this, any rows with missing data were dropped entirely.

3.2 Model Creation

The 'r8.6_k10_fct2' variable in the dataset is a two-level variable based on the K10 score, where the range 0–11 represents none/low/moderate and the range 12–40 represents high/very high [12]. Based on these K10 levels, we created binary targets: Low Distress (0–11) and High Distress (12–40). In this study, five single ML classifiers – the LR, SVM, ANN, NB and DT – and three ensembles – the RF, Adaptive Boosting (AD) and Gradient Boosting (GB) – were used for modelling and prediction analysis. The LR, SVM, ANN, NB, DT and RF

were selected based on their demonstrated success in related studies [6,13,25,29,34,35], whereas AB and GB were selected because they performed well in our previous work [4].

The LR classifier [24] is a generalised linear model [17,26]. Generalised linear models overcome limitations of linear models – including the use of dependent variables that are continuous and normally distributed, which are not always desirable – by using non-normal dependent variables [10,11]. In LR, the dependent variables can either be unordered or ordered polytomous, while the independent predictor variables can either be interval/ratio or dummy variables [24].

The SVM is a supervised learning model that learns from training data and performs classification on new data. It separates different classes by a hyperplane, and then maximises the separation distance as much as possible. Larger the margin, lower the error generated by the classifier [5].

The ANN is a feedforward neural network that uses supervised learning. This algorithm continually computes and updates all the weights in its network to minimise error. It consists of two phases: a feedforward phase where the training data is forwarded to the output layer; and the second phase, where the difference between this output and the desired target (the error) is backpropagated to update the weights of the network [32].

The DT classifier is based on Hunt's algorithm [18], and was developed by Quinlan [30]. It builds a tree-like decision model for classification and prediction, and is a useful explanatory tool for expressing the cause and effect chain [31]. It is typically used as a base classifier for ensemble models (e.g., RF, and AB).

NB is the simplest form of Bayesian network classifiers given the independence of each feature. Nevertheless, many applications have successfully implemented NB, and it is included among the top 10 data mining algorithms [19].

The RF is an ensemble of DT predictors where each tree is independently trained using a random vector. Error generalisation of RF depends on the strength of each individual tree and the correlation between them. This ensemble model is relatively robust to outliers and noise [2].

The AB ensemble algorithm iteratively combines multiple weak classifiers over several rounds. It starts with equal weights for all training data. When the training data points are misclassified, the weights of these data points are boosted, then a new classifier is created using the new unequal weights. This process is repeated for a set of classifiers [38].

GB is an ensemble of gradient boosted regression trees for the classification of dirty data, which produces a robust, competitive and interpretable algorithm for classification and regression. However, it uses only a single regression tree for binary classification [14].

As the dataset used in this study is relatively small (n = 2,010, and 1,985 after normalisation), the 10-fold cross-validation technique was applied. 10-fold cross-validation requires the dataset to be randomly partitioned into 10 equal subsets. 10 model building and test runs were then completed, each time utilising a different arrangement of 9 subsets for training and 1 subset for testing [25]. The

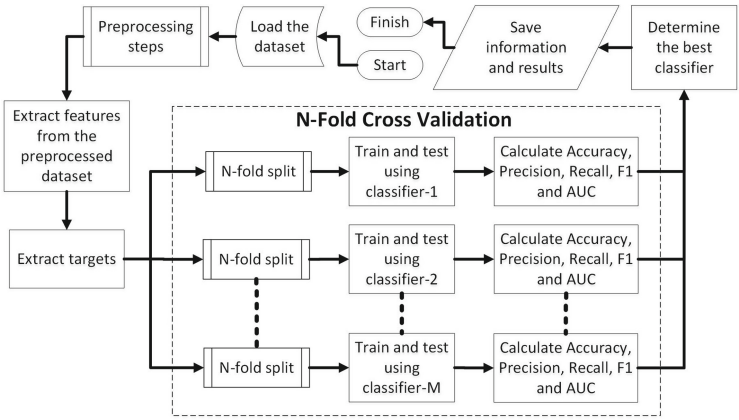


Fig. 1. An overview of the experiment workflow

entire experiment workflow is described in Fig. 1. All the code for the experiments was written and run on Google Colabs using Scikit Learn [27].

Five metrics, namely the accuracy, precision, recall, AUC, and F1 score (weighted average of precision and recall), were used for analysing the results of the proposed model, as well as for comparing them with the results obtained using other models from the literature. Accuracy was calculated by taking the number of correct predictions on the test set. Precision was calculated using Eq. 1, where tp is the number of true positives, and fp is the number of false positives [27]. Recall was calculated using Eq. 2, where fn is the number of false negatives [27]. Equation 3 is then used with precision and recall values to give the F1 score.

$$tp/(tp + fp) \tag{1}$$

$$tp/(tp + fn) \tag{2}$$

$$2 * (precision * recall)/(precision + recall) \tag{3}$$

3.3 Classification Analysis

Model validation is critical in ML training in order to prevent the over-fitting of a specific trained model. Hence, during the final phase of this study, the results of each classification model were compared to each other. As in the model creation phase, precision, recall, AUC and F1 scores were used as primary metrics to measure the performance of each classifiers.

4 Experiments and Results

All experiments were run using the 10-fold validation technique and averages were taken over 10 runs. Hyperparameters remained constant based on their default implementations.

Table 2 shows the average accuracy, precision, recall, F1 and AUC scores of each classifier. The results indicate that the LR model provided the best results. Its AUC score of 0.730 indicates that it made more correct than incorrect predictions. Similarly, with a recall score of 0.918, the LR model accurately predicted those positive cases. Therefore, despite the potential for improvement, we can conclude that our model can accurately predict psychological distress using ecological factors alone. Among other single classifiers, the ANN also performed well (accuracy of 0.807, better precision but a lower recall value than LR). The NB classifier had the best precision but lower accuracy – meaning that it made more similar mistakes than other classifiers. The DT and SVM, while still providing accuracies above 70%, seem less suitable for psychological distress prediction than the other models.

The ensemble models tested in our study (AB, GB, and RF) also generally performed well, with the RF providing slightly worse results than the others. Given that the GB ensemble uses a regression tree as its base classifier, the good results are both expected and obvious. As discussed above, LR, which is based on regression analysis, was the best classifier; however, the DT, which is the base classifier of AB and RF, performed only moderately. In other words, the ensemble models performed well despite their base classifiers performing only moderately. This means that better results can be achieved if we boost the weak classifiers continuously (as in AB) or bind some weak single classifiers (e.g., DT) in the RF.

Table 2. Results obtained with different classification methods

Classifier	Accuracy	Precision	Recall	F1	AUC
LR	0.811	0.835	0.918	0.875	0.730
AB	0.810	0.835	0.916	0.874	0.729
GB	0.810	0.837	0.912	0.873	0.732
ANN	0.807	0.840	0.904	0.871	0.734
RF	0.795	0.836	0.888	0.861	0.723
NB	0.778	0.856	0.830	0.843	0.739
DT	0.736	0.820	0.809	0.814	0.681
SVM	0.736	0.835	0.815	0.825	0.678

As the input dataset was mapped 1:1 to the feature layout of the model, it is possible that further manipulation could enhance the performance of the model. Such manipulation may involve adjusting weights based on bias, removing non-dominant features or implementing feature crosses that would provide a better depiction of the data in the given context. Additionally, the performance of the model could also be enhanced through tuning of the hyperparameters.

Our experiments utilised the MLPClassifier class (i.e., multilayer perceptron) of the Scikit Learn library to implement an ANN [27]. Related studies have

shown neural networks to successfully predict in areas of psychological distress [34]. Therefore, results may be further improved by re-implementing an ANN model, or implementing additional neural network models using specialised neural network frameworks such as TensorFlow Keras [8].

Further work on the actual ecological metrics included in the dataset would also be necessary to optimise the models. Removing metrics with little impact on K-10 scores, and adding further metrics with known positive impacts on K-10 scores would likely improve the model's scores. It is also important to note that the data sample used in this study was generated in the context of the COVID-19 pandemic, for the purpose of a study in that context. Therefore, we believe that a different data sample within the context of a more holistic, generalised view may also be beneficial.

5 Conclusion

In this paper, we proposed an ML-based model for psychological distress prediction using only ecological factors. Implementing eight classifiers using Scikit Learn [27], our LR classifier produced the best results, presenting an AUC of 0.73. Although below the 0.8 target, its accuracy of 0.811, precision of 0.835 and recall of 0.918 suggested that the model can accurately predict positive cases of psychological distress. Our results indicated that, although it is possible to create an ML model to predict psychological distress, the challenge lies in finding suitable ML model parameters and ecological features. Future work in this area would be to further analyse and tweak parameters to enhance the current models. Accuracy may also be improved by implementing alternative ecological factors as metrics in order to provide a greater holistic view.

Once an accurate model has been built, it can be used to bridge the gap in existing research in the literature, and also incorporated into real world software or mobile applications. This could include the integration with brain activity data [28], text sequence classification [34], or possibly with wearable devices to provide sleep, activity and heart rate information. With an enhanced model using metrics from multiple areas, some in real time, it will then be possible to provide the proactive approach required to effectively deal with this mental health crisis.

Acknowledgement. The first author would like to acknowledge financial support from a Research and Innovation Summer Research Internship Program scholarship awarded by the University of Newcastle, Australia.

References

1. Australian Institute of Health and Welfare: Mental health services in Australia: in brief 2019. AIHW (2019). https://doi.org/10.25816/5ec5bac5ed175
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
3. Brooks, B.L., et al.: Predicting psychological distress after pediatric concussion. J. Neurotrauma **36**(5), 679–685 (2019)

4. Budhi, G.S., Chiong, R., Pranata, I., Hu, Z.: Using machine learning to predict the sentiment of online reviews: a new framework for comparative analysis. Arch. Comput. Methods Eng. 1–24 (2021). https://doi.org/10.1007/s11831-020-09464-8
5. Campbell, C., Ying, Y.: Learning with Support Vector Machines. Morgan & Claypool (2011)
6. Chiong, R., Fan, Z., Hu, Z., Chiong, F.: Using an improved relative error support vector machine for body fat prediction. Comput. Methods Programs Biomed. **198**, 105749 (2021)
7. Chiong, R., Satia Budhi, G., Dhakal, S.: Combining sentiment lexicons and content-based features for depression detection. IEEE Intell. Syst. **36** (2021, in press)
8. Chollet, F., et al.: Keras (2015). https://keras.io
9. Cornelius, B.L.R., Groothoff, J.W., van der Klink, J.J.L., Brouwer, S.: The performance of the K10, K6 and GHQ-12 to screen for present state DSM-IV disorders among disability claimants. BMC Public Health **13**(1), 1–8 (2013)
10. Dobson, A., Barnett, A.: An Introduction to Generalized Linear Models, 3rd edn. CRC Press, Boca Raton (2008)
11. Dunteman, G., Ho, M.: Generalized linear models. In: An Introduction to Generalized Linear Models, pp. 2–6. SAGE Publications, Inc. (2011)
12. Every-Palmer, S., et al.: Psychological distress, anxiety, family violence, suicidality, and wellbeing in New Zealand during the COVID-19 lockdown: a cross-sectional study. PLOS ONE **15**(11), e0241658 (2020)
13. Fatima, I., Mukhtar, H., Ahmad, H.F., Rajpoot, K.: Analysis of user-generated content from online social communities to characterise and predict depression degree. J. Inf. Sci. **44**(5), 683–695 (2018)
14. Friedman, J.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
15. Furukawa, T.A., Kessler, R.C., Slade, T., Andrews, G.: The performance of the K6 and K10 screening scales for psychological distress in the Australian national survey of mental health and well-being. Psychol. Med. **33**(2), 357–362 (2003)
16. Galatzer-Levy, I.R., Karstoft, K.I., Statnikov, A., Shalev, A.Y.: Quantitative forecasting of PTSD from early trauma responses: a machine learning application. J. Psychiatr. Res. **59**, 68–76 (2014)
17. Hastie, T., Tibshirani, R.: Generalized Additive Models. Chapman and Hall/CRC, UK (1990)
18. Hunt, E., Marin, J., Stone, P.: Experiments in Induction. Academic Press, New York (1966)
19. Jiang, L., Li, C., Wang, S., Zhang, L.: Deep feature weighting for Naïve Bayes and its application to text classification. Eng. Appl. Artif. Intell. **52**, 26–39 (2016)
20. Kessler, R.C., et al.: Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol. Psychiatry **21**(10), 1366–1371 (2016)
21. Leightley, D., Williamson, V., Darby, J., Fear, N.T.: Identifying probable post-traumatic stress disorder: applying supervised machine learning to data from a UK military cohort. J. Ment. Health **28**(1), 34–41 (2019)
22. Loula, R., Monteiro, L.H.A.: A game theory-based model for predicting depression due to frustration in competitive environments. Comput. Math. Methods Med. **2020**, 3573267 (2020)
23. Patel, M.J., Khalaf, A., Aizenstein, H.J.: Studying depression using imaging and machine learning methods. NeuroImage Clin. **10**(C), 115–123 (2016)
24. Menard, S.: Logistic Regression: From Introductory to Advanced Concepts and Applications. SAGE, Los Angeles (2010)

25. Mor, N.S., Dardeck, K.L.: Quantitative forecasting of risk for PTSD using ecological factors: a deep learning application. J. Soc. Behav. Health Sci. **12**(1), 61–73 (2018)
26. Nelder, J.A., Wedderburn, R.W.: Generalized linear models. J. R. Stat. Soc. Ser. A (General) **135**(3), 370–384 (1972)
27. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
28. Portugal, L.C., et al.: Predicting anxiety from whole brain activity patterns to emotional faces in young adults: a machine learning approach. NeuroImage Clin. **23**, 101813 (2019)
29. Priya, A., Garg, S., Tigga, N.P.: Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Comput. Sci. **167**, 1258–1267 (2020)
30. Quinlan, J.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
31. Rokach, L., Maimon, O.: Data Mining with Decision Trees: Theory and Applications. World Scientific Publishing, Singapore (2007)
32. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science (1985)
33. Sau, A., Bhakta, I.: Screening of anxiety and depression among the seafarers using machine learning technology. Inform. Med. Unlocked **16**, 100149 (2019)
34. Trotzek, M., Koitka, S., Friedrich, C.M.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Trans. Knowl. Data Eng. **32**(3), 588–601 (2020)
35. Walsh, C.G., Ribeiro, J.D., Franklin, J.C.: Predicting risk of suicide attempts over time through machine learning. Clin. Psychol. Sci. **5**(3), 457–469 (2017)
36. Walsh, C.G., Ribeiro, J.D., Franklin, J.C.: Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. J. Child Psychol. Psychiatry **59**(12), 1261–1270 (2018)
37. World Health Organization: Other common mental disorders: Global health estimates. Geneva: World Health Organization, pp. 1–24 (2017)
38. Zhu, J., Zou, H., Rosset, S., Hastie, T.: Multi-class adaboost. Stat. Interface **2**, 349–360 (2009)