Department of Statistics,
University of Colombo

ICDS2021

Center for Data Science

Center for Data Science
University of Colombo, Sri Lanka

PROCEEDINGS OF THE INTERNATIONAL CONFERENCE IN DATA SCIENCE 2021

Data Science: Reshaping the Future

ICDS 2021

Data Science: Reshaping the Future

29th - 30th June 2021

Colombo, Sri Lanka

Proceedings
# of the
International Conference
in Data Science 2021
(ICDS2021)

29th - 30th June 2021

Colombo, Sri Lanka

ICDS2021 is organized by the Center for Data Science

jointly with the

Department of Statistics, University of Colombo

Publisher: Center for Data Science, University of Colombo

ISBN 978-624-5873-02-9

Proceedings of the International Conference in Data Science 2021
29th - 30th June 2021, Colombo, Sri Lanka

Editors:
Dr. Dilhari Attygalle,
Department of Statistics, University of Colombo

Dr. Rushan Abeygunawardana,
Department of Statistics, University of Colombo

Dr. H. A. S. G. Dharmarathne,
Department of Statistics, University of Colombo

Dr. K. A. D. Deshani,
Department of Statistics, University of Colombo

Date published: 29th June 2021

Published by:
Center for Data Science,
Department of Statistics, Faculty of Science, University of Colombo, Colombo 03, Sri Lanka

ISBN 978-624-5873-02-9

Responsibility for the contents of these articles rests squarely upon their respective author(s) and not upon the editors or the reviewers. Data presented and conclusions reached by the author(s) are for information only and are not for use without independent substantiating investigation on the part of the user.

Welcome Message by the Conference Chair and Director Center for Data Science

On behalf of the organizing committee, it is my great pleasure to welcome you to the first ever International Conference in Data Science (ICDS 2021) in Sri Lanka. This two-day virtual conference is organized by the Center for Data Science jointly with the Department of Statistics, Faculty of Science, University of Colombo to celebrate the 5th anniversary of the Center for Data Science, 20th anniversary of the Department of Statistics and as part of the Centenary celebration of the Faculty of Science, University of Colombo. Our theme for the conference "Data Science: Reshaping the Future" has been carefully chosen to mark these milestones.

We have an exciting program lined up for this virtual conference, with a keynote speech, two guest speeches and six technical sessions spanning Data Visualization to Deep Learning and AI, facilitating collaborations between academia, industries and students. On the second day of the conference, we have organized two workshops followed by the final round of the mini-hackathon where more than eighty undergraduate teams gave their best in solving a real life problem utilizing cutting edge data science methodologies. I hope you will have an educational, productive and enjoyable time at this very special conference.

My sincere gratitude goes to the invited speakers, presenters and co-authors, workshop attendees and resource personnel, and organizers of the mini-hackathon. I would like to thank our sponsors; strategic and knowledge partner – Creative Software and technical partner Altria for providing generous financial support.

To organize a conference of this magnitude in the midst of a pandemic is not an easy task. The organizing committee planned and executed all activities virtually. We are now rolling out the proceedings virtually as well. I would like to thank the Vice-Chancellor of the University of Colombo, Senior Professor Chandrika N. Wijeyaratne for her wisdom and guidance. Also I would like to thank Dean of the Faculty of Science, Senior Professor and Chair of Physics Upul Sonnadara for his continuous and invaluable support. A very special thank you goes to the Board of Management of the Center for Data Science for the encouragement and the blessings given. Last but not least, I must thank the Head of the Department of Statistics, Professor Chandima Tillakaratne and all the members of the staff for their untiring effort to make this conference a success.

Dr. Sameera Viswakula
Conference Chair/ ICDS 2021
Director/ Center for Data Science

Message by the Vice-Chancellor, University of Colombo

I convey my sincere congratulations on the occasion of the International Conference in Data Sciences (ICDS 2021) organized by the Centre for Data Sciences and the Department of Statistics in celebration of the centenary year of the Faculty of Science, University of Colombo (UoC).

Your chosen theme “Data Science: Reshaping the Future” is time appropriate and aptly addressed by the excellent mix of lecture topics, thematic workshops and mini hackathons for students. I salute the academic and administrative leads of the Faculty of Science for recognizing the emerging importance of Data Sciences, conceptualizing UoC’s strategic development of the field and your strong commitment to enhance the required human capital in Sri Lanka. Your hard work has paid dividends through the establishment of the state-of-the-art Centre, an enabling infrastructure and the vibrant academic programs for undergraduates and postgraduates. Your commitment extends further into the development of partnerships with the corporate sector and the industry.

The upscaling of pedagogy, research and scholarship in Data Sciences witnessed in the recent past has encouraged intramural and extramural multidisciplinary and multisector engagement, which is very progressive. The need for a fully-fledged higher educational facility in big data analysis and the related subjects was championed by visionaries and policy makers of our University. The current academic staff are aptly qualified with doctoral degrees from highly recognized universities to be excellent trainers, supervisors and public intellectuals with international recognition. I take this opportunity to thank the head and members of the Department of Statistics for your lively engagement in the holistic professional development of every graduate of the Faculty of Science.

I am confident that this international conference, held during unprecedented times of the COVID19 pandemic, will add greater value to the learning environment of the Faculty of Science.

I extend my warmest wishes to all the speakers, organizers and participants of this conference. I am certain that this virtual conference will yield fruitful and productive deliberations entwined with the establishment of strong academic networks, while preserving the best of health and safety.

May you have a fulfilling future in Data Sciences at the University of Colombo!

Senior Professor Chandrika N Wijeyaratne
Vice Chancellor, University of Colombo

Message by the Dean, Faculty of Science, University of Colombo

On behalf of the staff and students of the Faculty of Science, University of Colombo, I extend my warmest welcome to all the participants of the 1st International Conference on Data Science 2021, ‘Reshaping the Future’. This conference is organized by the Center for Data Science in collaboration with the Department of Statistics to mark the 20th anniversary of the establishment of a separate department for Statistics under the Faculty of Science, University of Colombo.

The hosting of the ICDS 2021 conference at the University of Colombo is a very special occasion for us as we are celebrating 100 years of excellence in science education this year. ICDS 2021 offers an excellent platform for the exchange of scientific and technical knowledge and information related to the emerging field of Data Science, across many application areas. With the wide expanse of conference tracks ranging from machine learning to ethics of data use and with experts coming together from the industry and academia, we expect the knowledge sharing to be at a high level at this conference.

A total of 6 separate sessions are being conducted during the 2 days of the conference, including an inaugural session with keynote addresses, presentation of technical papers and post conference workshops focused on machine learning and deep learning followed by a mini hackathon.

I would like to take this opportunity to thank all the presenters and their co-authors for contributing to the dissemination of their research findings and the participants for registering at the conference. On behalf of the Faculty of Science, I extend my sincere gratitude to the organizing committee for working under difficult conditions due to the covid-19 pandemic, to make this event a success.

Senior Professor Upul Sonnadara
Dean of the Faculty of Science, University of Colombo

Message by the Head, Department of Statistics, University of Colombo

The International Conference in Data Science 2021 (ICDS 2021) is a joint endeavor of the Center for Data Science (CDS) and the Department of Statistics of the University of Colombo. This is the first conference in Data Science organized in Sri Lanka. It is organized to commemorate the 100 years of excellence in Science at the University of Colombo, 20th anniversary of the Department of Statistics and 5th anniversary of the CDS. Being able to organize an international conference in the same year which it celebrates the 5th anniversary, is a great achievement of the CDS.

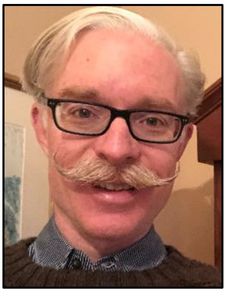
This conference provides a platform for data scientists to communicate the novel concepts in data science as well as the applications of data science techniques to solve real world problems. It also enhances the collaboration between academia and industry.

ICDS 2021 includes keynote and guest speeches, several contributed sessions, two workshops and a mini hackathon. CDS collaborates with the Acuity Knowledge Partners to conduct one of the workshops while collaborating with Creative Software and the Stat Circle, the student society associated with the Department of Statistics, to conduct the mini hackathon.

I take this opportunity to extend my sincere thanks to the chairperson and the organizing committee of the conference and wish all the best to the participants of the conference. I am confident that the ICDS 2021 will open doors to new developments and enhance the research in the discipline of Data Science. Furthermore, it will provide a platform to strengthen the ties between academia and the industry.

Professor Chandima Tilakaratne
Head, Department of Statistics, University of Colombo

Keynote Speaker



Jay Emerson is Director of Graduate Studies in the Department of Statistics and Data Science at Yale University. His academic work has included Bayesian change point analysis, statistics in sports, computational statistics and graphics, and environmental science. He is the lead statistician of the Yale Environmental Performance Index, and is working to apply similar rigor to the analysis of ESG metrics for applications in finance.

Guest Speaker

Bibhas Chakraborty is an Associate Professor and Ex-Director of the Centre for Quantitative Medicine at the Duke-National University of Singapore Medical School (Duke-NUS), as well as an Associate Professor of Statistics and Applied Probability at the National University of Singapore. He also holds an Adjunct Associate Professor position at the Department of Biostatistics and Bioinformatics at Duke University. Previously (2009-13), he was an Assistant Professor of Biostatistics at Columbia University. He completed his Ph.D. in Statistics from the University of Michigan, under the supervision of Prof. Susan A. Murphy in 2009. He is the recipient of the Calderone Research Prize for Junior Faculty from Columbia University's Mailman School of Public Health in 2011, and the Young Statistical Scientist Award from the International Indian Statistical Association (IISA) in 2017. His core areas of research include statistical reinforcement learning, precision medicine, dynamic treatment regimens, mobile/digital health, adaptive clinical trial designs, and a variety of applications in clinical and behavioral sciences. He wrote the first textbook on dynamic treatment regimens.

Guest Speaker

Umashanger Thayasivam, PhD, earned his doctoral degree for his research work on Mixture distribution at the University of Georgia. He is currently a Professor in Statistics & Data Science at the Rowan University New Jersey. At the university he lectures and trains both undergraduate and graduate students. He received the Teaching Wall of Fame award for his excellence in teaching in 2018. His interdisciplinary statistical research has spanned diverse areas including mixture distribution, robust estimation, high dimensional data analysis, statistical data mining, biomarker discovery, and educational data mining. He has an excessive experience with collaborating scientist engineers. He has several journal publications and numerous conference presentations. He has been PI/co-PI for several internal and external grants and collaborations, including the recent data mining project with Bristol Myers Squibb pharmaceutical company as well as the NIH grant on blood-based biomarkers for early-stage Alzheimer's disease. He also has extensive experience in mentoring student research in statistical/data mining research within the last ten years mentored more than 40 undergraduate students, many of which have presented their works at regional and national conferences. Several of his publications are with undergraduate students.

International Conference in Data Science 2021 (ICDS2021)

29th - 30th June 2021, Colombo, Sri Lanka

Conference Committee

Advisory Committee

- **Senior Professor Chandrika N. Wijeyaratne,**
  Vice-Chancellor, University of Colombo, Sri Lanka
- **Senior Professor Upul Sonnadara,**
  Dean, Faculty of Science, University of Colombo, Sri Lanka
- **Senior Professor M. R. Sooriyarachchi,**
  Department of Statistics, University of Colombo, Sri Lanka
- **Dr. Nimal Wickremasinghe,**
  Former Professor in Statistics, University of Colombo, Sri Lanka
- **Dr. Ruvan Weerasinghe,**
  Senior Lecturer, University of Colombo, School of Computing, University of Colombo, Sri Lanka

Members of the Board of Management - Center for Data Science:

- Prof. Upul Sonnadara - Dean/Faculty of Science (Chairman}
- Prof. C.D. Tilakaratne (Head/Department of Statistics)
- Dr. Sameera Viswakula (Director/Center for Data Science)
- Prof. S. Karunaratne (Representative from the Council)
- Mr. J.M.U.B. Jayasekara (Representative from the Council)
- Prof. Pavithra Kailasapathy (Representative from the Senate)
- Prof. Chamari Hettiarachchi (Representative from the Senate)
- Dr. Pemantha Lakraj (Representatives from the Department of Statistics)
- Dr. Rasika Jayatillake (Representatives from the Department of Statistics)
- Dr. Nadeeka Basnayake (Representatives from the Department of Statistics)
- Dr. Rohitha Wijewardena -Director, Altria Consulting (Pvt) Ltd
  (Representative of Industry experts)

- Dr. Rashmi Bomiriya, Managing Director - R S Metrics Asia Holdings & Chief Data Scientist, Remote Sensing Metrics Pvt Ltd (Representative of Industry experts)

Members of the Department of Statistics:

- Professor C. D. Tilakaratne – Head of the Department (Professor)
- Senior Professor M. R. Sooriyarachchi (Senior Professor)
- Dr. Dilhari Attygalle (Senior Lecturer)
- Dr. Rushan Abeygunawardana (Senior Lecturer)
- Dr. J. H. D. S. P. Tissera (Senior Lecturer)
- Dr. C. H. Magalla (Senior Lecturer)
- Mr. E. R. A. D. Bandara (Senior Lecturer)
- Dr. G. P. Lakraj (Senior Lecturer)
- Dr. R. V. Jayatillake (Senior Lecturer)
- Dr. S. D. Viswakula - Conference Chair and Director CDS (Senior Lecturer)
- Dr. A. A. Sunethra (Senior Lecturer)
- Dr. K. A. D. Deshani (Senior Lecturer)
- Dr. N. D. Basnayake (Senior Lecturer)
- Dr. H. A. S. G. Dharmarathne (Senior Lecturer)
- Ms. Lakmini K. N. Wijesekara (Lecturer)
- Mr. A. R. Munasinghe (Lecturer)
- Dr. I. T. Jayamanne (Lecturer)
- Ms. H. S. Karunarathna (Lecturer)
- Ms. D S Wickramarachchi (Lecturer)
- Ms. Avanthi Saumyamala (Lecturer)
- Ms. K. M. Manushi Hansani Siriwardana (Temporary Assistant Lecturer)
- Ms. S. Ishanka Randini Fernando (Temporary Assistant Lecturer)
- Mr. M. D. B. P. P. Badullahewage (Temporary Assistant Lecturer)
- Ms. K. D. O. Virajini (Temporary Assistant Lecturer)
- Ms. A. P. N. G. Adhikari (Temporary Assistant Lecturer)
- Ms. M. R. R. G. Maha Ranawaka (Temporary Assistant Lecturer)

- Ms. E. S. P. H. Rajadasa (Temporary Instructor)
- Ms. T. M. M. Jayaweera (Temporary Instructor)
- Ms. Lankadinee W. Rathuwadu (Temporary Instructor)
- Mr. A. A. K. T. Amarasinghe (Temporary Instructor)
- Ms. .G. C. J. Piyatilake (Temporary Instructor)
- Ms. Jayamini C. Liyanage (Temporary Instructor)
- Ms. S. D. Daluwatumulle (Temporary Instructor)
- Ms. N.G.N. Abirami (Temporary Instructor)
- Mr. H. K. T. Nanayakkara (Management Assistant)
- Mr. N.D Suduwella (Assistant Network Manager)
- Ms. R.M.N.E.K. Rathnayake (Management Assistant)
- Ms. R. A. K. Kithmini (Lab Attendant)
- Mr. W. D. M. C. Withanage (Support Staff)

Panel of Reviewers

All abstracts included in the Proceedings of the International Conference in Data Science 2021 have been independently reviewed through a double-blind process. The Advisory Committee and the Staff of the Department of Statistics would like to thank the following reviewers for their valuable services.

- **Dr. Rushan Abeygunawardana**
  Senior Lecturer, Department of Statistics, University of Colombo
- **Dr. Radheeka Abeyweera**
  Manager, Research and Development in Marketing Science, Facebook Inc.
- **Prof. Dhammika Amaratunga**
  Independent Researcher
- **Dr. Anuradha Ariyaratne**
  Senior Lecturer, Department of Computer Science, University of Sri Jayewardenepura
- **Dr. Dilhari Attygalle**
  Senior Lecturer, Department of Statistics, University of Colombo
- **Mr. E.R.A.D. Bandara**
  Senior Lecturer, Department of Statistics, University of Colombo
- **Dr. N.D. Basnayake**
  Senior Lecturer, Department of Statistics, University of Colombo
- **Dr. Rashmi Bomiriya**
  Managing Director, Asia and Chief Data Scientist, RS Metrics Holdings
- **Dr. Vasana Chandrasekara**
  Senior Lecturer, Department of Statistics & Computer Science, University of Kelaniya
- **Prof. Sanjay Chaudhuri**
  Associate Professor, Department of Statistics and Applied Probability, National University of Singapore.
- **Dr. Mahasen Dehideniya**
  Senior Lecturer, Department of Statistics & Computer Science, University of Peradeniya
- **Dr. K.A.D. Deshani**
  Senior Lecturer, Department of Statistics, University of Colombo
- **Dr. Neluka Devpura**
  Senior Lecturer, Department of Statistics, University of Sri Jayewardenepura
- **Dr. H. A. S. G. Dharmarathne**
  Senior Lecturer, Department of Statistics, University of Colombo

- **Dr. Malathi Imiyage Dona**
  Research Officer, Cardiac Cellular Systems, Baker Heart and Diabetes Institute, Australia
- **Dr. Nusrath Hameed**
  Senior Lecturer, Department of Computer Science, University of Ruhuna
- **Dr. Gayan Hettiarachchi**
  Chief Scientist, OpenDNA Inc. Japan. Executive External Advisor, Sierra Consulting Inc. USA. Visiting academic staff, Osaka University and Rutgers University.
- **Dr. Isuru Hewapathirana**
  Senior Lecturer, Software Engineering Teaching Unit, Faculty of Science, University of Kelaniya
- **Dr. I.T. Jayamanne**
  Lecturer, Department of Statistics, University of Colombo
- **Ms. Mandu JayasundeDera**
  Research Scientist, Facebook Inc.
- **Dr. R.V. Jayatillake**
  Senior Lecturer, Department of Statistics, University of Colombo
- **Dr. J.A. Jeewani**
  Senior Lecturer, Department of Computer Science, University of Ruhuna
- **Dr. Kasun Karunanayaka**
  Senior Lecturer, University of Colombo School of Computing
- **Dr. Prasanna Karunanayake**
  Assistant Professor ,Department of Radiology, Department of Public Health Sciences, Department of Neural and Behavioral Sciences, Penn State Neuroscience Institute, Penn State University, USA
- **Ms. G.H.S. Karunarathna**
  Lecturer, Department of Statistics, University of Colombo
- **Dr. Chamath Keppitiyagama**
  Senior Lecturer, University of Colombo School of Computing
- **Dr. Chinthaka Kuruwita**
  Associate Professor of Statistics, Hamilton College, New York, USA
- **Dr. G.P Lakraj**
  Senior Lecturer, Department of Statistics, University of Colombo
- **Dr. C. H. Magalla**
  Senior Lecturer, Department of Statistics, University of Colombo

- **Mr. Rajith Munasinghe**
  Lecturer, Department of Statistics, University of Colombo
- **Dr. Hasanthi Pathberiya**
  Senior Lecturer, Department of Statistics, University of Sri Jayewardenepura
- **Dr. Dilina Perera**
  Senior Lecturer, Department of Physics, University of Colombo
- **Mr. Shanaka Perera**
  PhD candidate, Department of Computer science, University of Warwick.
- **Dr. Charith Peris**
  Research Scientist, Alexa AI at Amazon
- **Dr. S. P. Pitigala**
  Senior Lecturer, Department of Statistics & Computer Science, University of Kelaniya
- **Dr. Chathura Rajapakse**
  Senior Lecturer, Department of Industrial Management, University of Kelaniya
- **Dr. Nishath Rajiv**
  Postdoctoral Researcher, Prairie View A&M University, USA
- **Ms. Avanthi Saumyamala**
  Lecturer, Department of Statistics, University of Colombo
- **Mr. Nuwan Senaratna**
  Independent Consultant
- **Ms. Devini Senaratna**
  Data Scientist, Booking.com
- **Dr. Sadun Malpriya Silva**
  Biostatistician, NHMRC Clinical Trials Centre, University of Sydney, Australia
- **Dr. A.A. Sunethra**
  Senior Lecturer, Department of Statistics, University of Colombo
- **Dr. Priyanga Talagala**
  Senior Lecturer, Department of Computational Mathematics, University of Moratuwa
- **Dr. Thiyanga Talagala**
  Senior Lecturer, Department of Statistics, University of Sri Jayewardenepura
- **Prof. Umashanger Thayasivam**
  Professor of Statistics and Data Science, Department of Mathematics, Rowan University, USA

- **Dr. Pushpike Thilakarathne**
  Statistician, Janssen, Pharmaceutical Companies of Johnson and Johnson
- **Prof. C.D.T. Tilakarathne**
  Professor, Department of Statistics, University of Colombo
- **Dr. J.H.D.S.P.Tissera**
  Senior Lecturer, Department of Statistics, University of Colombo
- **Dr. Hakim. A. Usoof**
  Senior Lecturer, Department of Statistics & Computer Science, University of Peradeniya
- **Dr. S.D. Viswakula**
  Senior Lecturer, Department of Statistics, University of Colombo
- **Dr. Ruvan Weerasinghe**
  Senior Lecturer, University of Colombo School of Computing
- **Mr. Viraj Welgama**
  Senior Lecturer, University of Colombo School of Computing
- **Dr. Dilani Wickramaarachchi**
  Senior Lecturer, Department of Industrial Management, University of Kelaniya
- **Ms. D. S. Wickramarchchi**
  Lecturer, Department of Statistics, University of Colombo
- **Dr. Shanika Wickramasuriya**
  Lecturer, Department of Statistics, University of Auckland, Australia
- **Ms. Lakmini K. N. Wijesekara**
  Lecturer, Department of Statistics, University of Colombo
- **Ms. Rupika Wijesinghe**
  Senior Lecturer, University of Colombo School of Computing
- **Mr. Hiran Wijesinghe**
  Assistant Director IT, Sri Lanka Tea Board

Table of Contents

Title and Author/s **Page No:**

Data Science: From 2015 to 2021 and Beyond **(Keynote speech)** 1
John W. Emerson ("Jay")

Data Science in Mobile Health **(Guest speech)** 2
Bibhas Chakraborty

Future (Challenges) of Data Science- Quantum Computing **(Guest speech)** 3
Umashanger Thayasivam

Technical Sessions

Predicting Career Satisfaction of IT Employees: A Case Study on Software Developers 5
Nimasha Arambepola, Lankeshwara Munasinghe

Improving Time Series Anomaly Detection Using an Ensemble Feature Model 6
H. Asela

A Case Study in Financial Fraud Detection using Big Data Analytics 7
W. P. A. Boteju, I. U. Hewapathirana

Prediction of Contact Maps and Family of Proteins using Deep Learning 8
N. H. Charuka , W. A. M. Madhavi

A Study on Vehicle Emission Test in Sri Lanka 9
K. S. P. Fernando, R. A. B. Abeygunawardana

Impact of the dollar rate in prediction of the Colombo Stock Exchange Performance using Machine Learning Techniques 10
Sachith Fernando, N. A. I. Supasan, Ranjan Dissanayake

A Machine Learning Algorithm to Validate Present Data with Historical Data 11
S. R. N. Gunaratne, K. A. D. Deshani, S. Srikathirkamanathan

Title and Author/s **Page No:**

Identifying the Factors Associated with Infant Mortality in Galle MOH Areas 12
M. Gunawardhana, R. V. Jayatillake, K.T. Samararathna

Seismicity And Stress Patterns Along The Central Tunisia 13
Makrem Harzali, Abdelmalak Mohamed Mansour

Spatial Changes and Relationship Analysis of NDVI and LST Using Satellite Image Datasets in Pottuvil Divisional Secretariat Division, Ampara 14
M. L. M. Hicmathulla, A. W. F. Nafla

Gamified Feedback and Job Satisfaction of Generation Z Software Professional in Post-COVID-19 Environment 15
H. D. Jayathilake, N. P. Barupala, J. A. Jayathilake, H. M. Jayathilake

A Deep Learning Approach to Hate Speech Detection in Sinhala 16
B. A. S. S. B. Jayawardhana, M. R. D. S. G. Punchihewa

Quantitative Evaluation of Equity Research for Accurate Investment Decisions in Future 17
D. R. S. Kalubowila, R. Fernando, S. D. Viswakula

Improving Sinhala Language Modelling through Deep Learning 18
D. S. Kulatunga, A. R. Weerasinghe, E. R. A. D. Bandara

Forecasting the Directional Movement and Value of S & P SL 20 Index Using News and Macro-Economic Indicators 19
M. P. Maha Arachchi, G. P. Lakraj, I. Wijesekara, P. Wijesiriwardana

Multi Target Regression for Click Through Rate Prediction 20
Rajitha Manellanga

Investigating the Chest Imaging Features of Covid-19 Patients using Deep Neural Networks 21
A. B. Meepaganithage, M. G. N. A. S. Fernando

Topic modeling in Youtube Data 22
A. C. Nanayakkara, G. A. D. M. Tennakoon

Title and Author/s **Page No:**

Predictive Analytics for Machinery Maintenance in the Production of Quality Garments 23
U.M.M.P.K.Nawarathne, Vijani Piyawardana, Jesuthsan Alosius

Study on Part-Time Employment of Science Undergraduates, University of Colombo 24
K. G. J. Nishadini, R. A. B. Abeygunawardana

Machine Learning Approaches to Forecast Inflation in Sri Lanka 25
W. P. I. S. Pathirana, C. D. Thilakaratne

A Comparative Study on Artificial Neural Network Techniques for Short-term Rainfall Forecasting in Colombo, Sri Lanka 26
G. C. J. Piyatilake, K. A. D. Deshani

Game Outcome Predictor and Game Plan Generator for Twenty20 Cricket 27
L. C. P. Pussella, R. M. Silva, W. C. P. Egodawatta

Machine Learning Approach to Detect Tea Plant Diseases in Sri Lanka 28
S. N. K. Rajakaruna, Rangana Jayashanka, H. A. S. G. Dharmarathne

Question Answering for Sinhala Language Using Deep NN Architectures 29
M. W. K. S. Rasantha, Dr. A. R. Weerasinghe, Dr. J. H. D. S. P. Tissera

Algorithm to Detect Doctored Images 30
R. M. D. S. Rathnayaka, S. D. Viswakula, T. A. M. C. Thantriwatte

Developing the Modelling Framework of an Intelligent News Recommendation Engine for Financial Analysts 31
H. M. P. P. K. H. Samarasekara, W. M. N. D. Basnayake, D. Hureekaduwa, B. Weralupitiya

Improving Bank-Customer Efficiency through Automatic Speech Recognition 32
L. D. P. S. Senaratne, H. N. D. Thilini, C. H. Magalla

Vehicle Detection In Satellite Images 33
K. Sutharsan, S. Gamika, Dr. G. P. Lakraj

Title and Author/s	**Page No:**
Linkage Comparison on Agglomerative Clustering for Online Retail Dataset **R. M. B. P. M. Uduweriya, N. A. D. N. Napagoda**	34
Building an OCR Tool with a Better Text Recognition for Financial Domain **B. L. A. S. Upekshika, S. D. Viswakula, M. F. S. Ahamed, F. L. Muthukumarasamy**	35
Spotting Railway Signs to Build Smart Decision Support Tools in Railway Management Systems **P. T. Weerasinghe, M. Shaheer, R. Rumalshan, P. Gunathilake, E. Dayarathne**	36
A Case Study of Creditworthiness Prediction at the Loan Approval **B. N. Weralupitiya and R. V. Jayatillake**	37
Quantifying the Heterogeneous Close Contact Patterns in Sri Lanka **W. A. P. S. Wickramarachchi, L. Munasinghe**	38
Workflow Composition with Predictive Modelling **C. R. Wijesinghe, A. R. Weerasinghe**	39
Developing a Model to Decrease Call Volume Cost of a Telecommunication Company by Analysing Time-to-Failure since the Installment of the SetupBox **K. A. C. N. Wijewardana, H. A. S. G. Dharmarathne,** **P. A. D. L. Samarakoon**	40

Data Science: From 2015 to 2021 and Beyond

John W. Emerson ("Jay")

Department of Statistics and Data Science, Yale University

This talk offers commentary and perspective on the evolution of Data Science through personal examples from Yale and around the world. Some observations I made in 2015 – at a Keynote Address right here in Colombo! – were accurate and, I hope, helpful. But I didn't place sufficient emphasis on the importance of collaboration. And I didn't predict the explosion of interest and activity in Data Science that we've witnessed in the last 6 years. I will correct those mistakes today in this non-technical talk that I hope will be accessible and enjoyable for everyone in attendance, whether in-person or remote. And unlike my 2015 address, I will not present code, run simulations, or consider topics in high-performance computing. My primary example will be the collaborative Environmental Performance Index project, ranking countries on environmental health and ecosystem vitality. It is a biennial analyses with roots going back to the 2000 Environmental Sustainability Index. I will leave more technical topics in Data Science to other speakers, your technical sessions, and conference workshops.

Data Science in Mobile Health

Bibhas Chakraborty

Centre for Quantitative Medicine and Program in Health Services and Systems Research,
Duke-NUS Medical School

Mobile health (mHealth) is a modern field in health sciences. mHealth interventions, broadly defined as health and behavioural interventions delivered via mobile and/or wearable devices, offer a powerful vehicle to improve health outcomes over a broad spectrum of the population in today's technology-enabled world. In recent years, healthcare providers and governments are excited about the opportunities mHealth offers in terms of improving efficiency and reducing cost of healthcare systems. With the increasing volume, quality and accessibility of mHealth data, statistics/data science has a key role to play in the development of data-driven, personalized, just-in-time adaptive interventions (JITAIs), and thereby in advancing human health. JITAIs often consist of sequences of text-messages or other prompts that aim to adapt to an individual's time-varying context (e.g., location, weather, past behaviour, past treatments, evolving health status, etc.), and thus optimizing them is a multi-stage decision problem.

In this talk, we will present an innovative experimental design called the micro-randomized trial (MRT), the current gold-standard data source for developing optimal JITAIs. This design involves sequential, within-individual randomizations, and aims to estimate proximal causal effects of "push"-type interventions that constitute JITAIs, e.g., motivational text-messages to promote physical activity or other healthy behaviours. JITAIs are usually learned via offline longitudinal data analysis after the MRTs are completed. While this is a sound randomization-based approach, trial participants do not get the benefit of receiving the optimized JITAIs. So, in order to make the MRT design more adaptive and thereby to deliver better interventions to the trial participants, we will consider contextual multi-armed bandit (MAB) type online reinforcement learning algorithms from the field of artificial intelligence, wherein a system must repeatedly choose among a set of actions (interventions) in an effort to maximize a reward (outcome). We will focus on a popular MAB algorithm called Thompson Sampling (TS) – deeply grounded in Bayesian statistics, particularly useful in case of small amounts of data, and often operationalized via a regression model for the reward. We will illustrate these ideas by discussing an mHealth case study in detail, and presenting real data analysis.

Future (Challenges) of Data Science- Quantum Computing

Umashanger Thayasivam

Department of Mathematics, Rowan University

Every day we create more than 2.5 quintillion data bytes of data, and this number is expected to grow to 3.5 quintillion data by 2025 especially with the rise of IoT (Internet of Things) and 5G capabilities. Data Science and artificial intelligence (AI) are some of the ways to help manage and analyze data for competitive advantage, however continued revolution and the desire for profound intuitions may make data more complex for organizations to collect and investigate. Quantum computing has the potential to be the most disruptive technology of the 21st century. This is a different form of computation that builds from quantum mechanics, and it promises to solve problems we cannot solve with classic computers.

Recently, Google announced that it had achieved quantum supremacy with its "Sycamore" quantum computer that can solve complex algorithms unsolvable by any other computer today. This milestone raises fundamental questions about how quantum computing can be used and how it will affect initiatives in the digital era. As classical binary computing reaches its performance limits, quantum computing is becoming one of the fastest-growing digital trends and is predicted to be the solution for the future's big data challenges. Though quantum computing is still just on the horizon, the U.S. plans to invest more than $1.2 billion toward quantum information over the next 10 years in a race to build the world's best quantum technology. Quantum computers will perform incredibly complex calculations in a couple of seconds when it would take a traditional computer a few thousand years to do the same. Google's Sycamore was able to perform a calculation in 200 seconds that would have taken today's fastest supercomputer 10,000 years to settle.

The future of data science will be impacted by quantum computing and some radical transformations in data science are necessary to develop quantum computing algorithms. Scientists hope quantum computing can help pave the way for the future of data science.

ICDS2021 - Technical Sessions

Predicting Career Satisfaction of IT Employees: A Case Study on Software Developers

Nimasha Arambepola, Lankeshwara Munasinghe

Software Engineering Teaching Unit, Faculty of Science, University of Kelaniya, Sri Lanka

Employees change their careers due to various reasons. For example, salary, changing career goals, working environment and job satisfaction are a few of them. Especially, the dynamic nature of the IT industry such as the diverse array of technologies, languages, frameworks, and platforms often cause software developers to change their career. Thus, the career satisfaction of IT employees highly depends on those factors. This study focuses on machine learning (ML) techniques for predicting the career satisfaction of software developers with the features extracted from employee job profiles. We observed that a handful of researchers have analysed the effectiveness of ML on predicting career satisfaction. However, those are not specifically for software developers. The present study analyses and compares the performance of popular supervised ML methods in predicting carrier satisfaction of software developers. Those ML methods are Random Forest (RF), Support Vector Machine (SVM) and Neural Network (NN). Stack Overflow developer survey published in 2018 was used as the benchmark dataset for our experimental evaluation. The ML models were trained to predict three levels of career satisfaction; Satisfied, Dissatisfied and Neither satisfied nor dissatisfied. When training the model, 76 features were selected out of 129, considering the RF feature importance. For example, the developer's salary, hours on the computer, years of coding, education type and company size are the most prominent features in our dataset. ML models were trained using 80% of the dataset and the remaining 20% used for testing. On average, the results show that the above ML models can predict the career satisfaction of software developers with an accuracy of around 80%. RF shows the highest accuracy which is 82% while SVM and NN show 79%.

Keywords: Career satisfaction, Classification models, Data mining, Machine learning

Improving Time Series Anomaly Detection Using an Ensemble Feature Model

H. Asela

Department of Electrical and Information Engineering, University of Ruhuna, Galle, Sri Lanka

Anomaly detection is a field of study in data mining and machine learning that involves identifying data, events, or observations deviating from their normal conduct. Anomaly detection in time series is used to monitor critical tasks in a variety of industries. These systems can be modelled using supervised learning models, unsupervised learning models, hybrid approaches, time series analysis, and statistical approaches. Unsupervised algorithms and ensemble models are used in many anomaly detection approaches since the data is mostly unlabeled and highly imbalanced. Also, the performance of these algorithms depends heavily on the feature set, as these features help to identify complex patterns in time series data. This research study focuses on integrating multiple approaches to build a robust feature set to improve anomaly detection performance in time series. The proposed feature set has 3 diverse categories of features. These features are generated from time series decomposition components, statistical metrics, and anomaly scores generated from multiple unsupervised machine learning algorithms. The proposed feature set is tested using the Isolation Forest unsupervised machine learning algorithm on Yahoo! S5 dataset with tagged anomaly points. The dataset contains 572,966 time series data instances in 4 data classes. F1 score, false alarm rate, and true positive rate are all used as evaluation metrics due to the imbalanced datasets. The proposed feature model showed an improvement in anomaly detection performance by increasing the F1 score by 1% - 62% and decreasing the false alarm rate by 1% - 38% across multiple data classes in the dataset. Furthermore, time series decomposition component features were found to be the most contributing set of features with a significant increase of true positive rate.

Keywords: Anomaly detection, Feature model, Isolation forest, Time series decomposition, Statistical metrics, Anomaly scores

A Case Study in Financial Fraud Detection using Big Data Analytics

W. P. A. Boteju, I. U. Hewapathirana

Software Engineering Teaching Unit, Faculty of Science, University of Kelaniya, Sri Lanka

The financial industry is currently undergoing digital transformations across products, services and business models. This digitization is aimed at automating most of the manual financial transactions and other relevant services. Therefore, spotting fraud in financial transactions has become an important priority for all financial institutes. With the advances in modern technology and global communication, fraud has increased significantly, causing great damages. The focus of this paper is to experiment different approaches for detecting fraudulent activities in a real-world dataset of financial payment transactions. The dataset is obtained from Kaggle and consists of 6 million transaction records and 10 features with the transaction label as 'fraudulent' or 'non-fraudulent'. These features are investigated using exploratory data analysis and only 6 are retained for the experiment such as payment-type, account-balance, transaction-amount etc. Two supervised machine learning algorithms, the random forest and the support vector classifier are employed for detecting fraudulent transactions. The dataset is large and requires high computational power to process and train machine learning algorithms. Furthermore, another challenge is the highly imbalanced distribution between fraudulent (0.1%) and the non-fraudulent (99.9%) classes. The goal of this research is to solve both these issues. In order to handle class imbalance, the effect of oversampling the minority class data using the synthetic minority oversampling technique (SMOTE), and undersampling the majority class using random undersampling are investigated. Computational efficiency is achieved through the Apache Spark implementation, which provides distributed processing for big data workloads. The best performance is obtained using the random forest algorithm on the oversampled dataset with an accuracy of 99.95%, F1-score of 0.9994, recall of 0.9994, Geometric mean of 99.94% and a model training time of 13.9 minutes. This paper provides valuable insights on dealing with large scaled highly imbalanced big datasets for predicting financial frauds and generating alerts.

Keywords: Financial Fraud Detection, Big Data Analytics, Apache Spark, SMOTE, Ensemble Learning Methods

Prediction of Contact Maps and Family of Proteins using Deep Learning

N. H. Charuka , W. A. M. Madhavi

Department of Physics, University of Colombo, Colombo, Sri Lanka

Proteins are large biological molecules present in all kinds of living things and are essential for their functioning. Even though there are many types of proteins, the basic building blocks of proteins are limited to a few amino acids. The process of finding the 3D structure of a protein when the amino acid sequence is given, is called the protein structure prediction problem. Protein contact map is a binary 2D matrix which indicates the distances between each pair of amino acid residues. Although folded proteins are in 3D space and the protein contact maps are 2D, contact maps can be used as representations of protein structures. In this study, a simple and intuitive neural network architecture is proposed to predict protein contact maps. The feasibility of this architecture was tested against a subset of the ProteinNet dataset. First layer of the neural network was an embedding layer. Then, there were 1D convolution and max pooling layers to reduce input data to a single vector in the middle of the neural network (encoding). Then this single vector was up sampled with 2D transpose convolution layers to produce 2D output data (decoding). By this method, the transformation from 1D to 2D which is required in the middle of the neural network is eliminated. Therefore, this method provides simplicity and intuition without complicated transformations. To predict the protein family, dense layers were added to the encoder part of the neural network, and it was tested for a subset of the PFam dataset. After training the neural network, the predicted contact maps were achieved with an accuracy of 97.45% and a precision of 90.29% for the amino acid sequences shorter than 128. This is comparable with the contact map predictions using MULTICOM-NOVEL, RaptorX-Contact, CONSIP2 methods that have given less than 90% precision values. For the PFam dataset, family prediction accuracy was 94.32%.

Keywords: Protein contact maps, Deep learning, Machine learning, Protein family prediction

A Study on Vehicle Emission Test in Sri Lanka

K. S. P. Fernando, R. A. B. Abeygunawardana

Department of Statistics, University of Colombo, Colombo, Sri Lanka

This research was carried out under two main objectives, identifying the most effective factors on Vehicle Emission Test (VET) and building a predictive model for classifying petrol and diesel vehicles into test status 'pass' and 'fail' was performed under the project. As there are two different vehicle emission testing methods used for petrol and diesel vehicles, this research considers two datasets which are known as "petrol dataset" and "diesel dataset". Both datasets which were procured from one of the pioneering organizations in air quality management and eco solutions in Sri Lanka. Since these two datasets are different from each other, the analysis was performed separately. The first objective was achieved through descriptive analysis and advanced analysis by using several machine learning techniques such as random forest, XGBoost and AdaBoost. The most significant factors for VET and XGBoost model were chosen to be the optimal model for both petrol and diesel vehicles. The optimal model was chosen by comparing misclassification errors in each model where XGBoost resulted in minimum misclassification errors for both petrol and diesel dataset and those values are 0.0023 and 0.0935 respectively. According to the results, acceleration CO, IdleCO, IdleHC and vehicle age were the main factors associated with the test result status of petrol vehicle emission. With regard to diesel vehicle emission tests, average opacity, number of cylinders, vehicle age and vehicle class were the associating factors. Using the information obtained from the research, a mobile app was created using android studio to provide a classification for VET results, pass or fail.

Keywords: Vehicle Emission Test, misclassification error, Opacity

Impact of the dollar rate in prediction of the Colombo Stock Exchange Performance using Machine Learning Techniques

Sachith Fernando, N. A. I. Supasan, Ranjan Dissanayake

Department of Physical Sciences, Faculty of Applied Sciences, Rajarata University of Sri Lanka

It is important to predict stock market performance to make future decisions on investments. Use of machine learning (ML) techniques in assessing the impacts of dollar rates on Colombo Stock Exchange (CSE) performance is pertinent and novel in Sri Lanka. The objective of this study was to examine the impact of dollar rate in predicting daily movements of the CSE using two different ML techniques namely multiple linear regression, and neural network. A total of 4,446 daily index values of the CSE and variables including high price, low price, open price, volume, and dollar rate were collected from 03/01/2000 to 31/12/2018 for the analysis. Of these data, 80% were used to build the models while 20% were used to validate the built models. The multiple linear regression model revealed that high price and the dollar rate have a significant impact on the prediction of CSE's closing price ($p<0.05$) with an accuracy of 97.27% (Mean Absolute Percentage Error (MAPE) = 2.73). The neural network model used Long Short-Term Memory (LSTM) algorithm to predict the closing price index of the CSE. Its accuracy fitted with and without the dollar rate were 98.32% (MAPE = 1.68) and 98.14% (MAPE = 1.86), respectively. The neural network model trained with the dollar rate showed a higher prediction accuracy than the model without the dollar rate. In conclusion, both the multiple linear regression model and neural network LSTM algorithm showed the significance of dollar rate on CSE performance prediction and the prediction accuracy of CSE performance was higher for neural network LSTM algorithm compared to multiple linear regression model.

Keywords: Colombo Stock Exchange, Mean Absolute Percentage Error, Multiple Linear Regression, Neural Network, Long Short-Term Memory

A Machine Learning Algorithm to Validate Present Data with Historical Data

S. R. N. Gunaratne[1], K. A. D. Deshani[1], S. Srikathirkamanathan[2]

[1]Department of Statistics, University of Colombo, Colombo, Sri Lanka

[2]Specialized Solutions, Acuity Knowledge Partners, Colombo, Sri Lanka

Business decisions made based on low-quality data affect companies negatively. This study aimed to formulate an algorithm based on machine learning (ML) techniques to validate client-sent data from multiple domains using its historical data. Hence, a validation method using prediction intervals with ML models was implemented. The proposed algorithm was applied to two datasets, scraped from two websites; a sports item and an airline ticket dataset. The latest scrape was kept as the new dataset to be validated and the rest of the scrapes were considered as historical data that was used to implement the models. Here, the new data follows a similar pattern as of its historical data. A response variable was then selected so that it is predicted by all the other independent variables in the dataset. 80% of historical data was then used for model training and 20% for model testing. Three ML techniques were used to model the data; Ordinary Least Squares, Random Forest and a Deep Neural Network. The model resulting in the lowest MAE was considered as the best model to represent the historical data and the technique linked with the best model was used to validate the newly scraped data using prediction intervals. Finally, if an observed value does not fall under the calculated prediction interval, that particular data point is not described accurately by its independent variables, and hence will be flagged as 'suspicious'. Results showed that the Random Forest model was the best for both datasets. After performing data validation using the proposed algorithm, 92% of the records were flagged as valid for one dataset and 49% for the other. The study reveals the possibility and ease of using ML in the area of data validation instead of the cumbersome techniques currently used where validation methods have to be customized from dataset to dataset.

Keywords: Machine Learning, Deep Learning, Data Validation, Prediction Intervals, Web Scraping

Identifying the Factors Associated with Infant Mortality in Galle MOH Areas

M. Gunawardhana[1], R. V. Jayatillake[1], K.T. Samararathna[2]

[1]Department of Statistics, University of Colombo, Colombo, Sri Lanka

[2]Department of Health, Ministry of Health, Colombo, Sri Lanka

According to the World Health Organization, infant mortality rate is the probability of a child born in a specific year or period dying before reaching the age of one. Although factors associated with infant mortality in Sri Lanka have been discussed in research, those were mainly done only with a descriptive approach and factor identification for separate districts is undiscovered. So, the importance of this research is to identify the demographic and health factors associated with infant mortality in Galle district, with an emphasis on using machine learning techniques. The dataset consisted of 240 observations including 115 infant mortalities and 125 healthy infants and 21 variables (available in pregnancy card). Since this was a case-control study balanced numbers were selected from cases (deaths) and controls (live past 1 year). The research was carried under two objectives. The first objective was to identify the factors associated with infant mortality. Classification techniques such as logistic regression, support vector machine, random forest (RF), XGBoost and artificial neural network were used since the response variable is categorical (dead, healthy). RF gave the highest accuracy of 93.72% and an AUC of 0.9956. To further investigate infant mortality, the second objective of identifying the factors associated with the number of days lived before an infant's mortality was carried out. Here the response variable is discrete, not continuous, and is limited to non-negative values. Therefore, several machine learning techniques regarding count models such as Poisson decision tree where the tree was pruned using minimum cross-validation error and Generalized Boosted Regression with distribution Poisson were used. Finally, the research concluded that from the importance plot of RF and common factors given by the importance plots of each model on the second objective, birth weight, maturity at birth and BMI of the mother are the highly associated factors for infant mortality in Galle district.

Keywords: infant mortality, case-control study, classification, machine learning, count models

Seismicity And Stress Patterns Along The Central Tunisia

Makrem Harzali[1,] Abdelmalak Mohamed Mansour[2,3]

[1]Water, Energy and Environment Laboratory, Engineering National School of Sfax (ENIS), University of Sfax, Sfax, Tunisia.

[2]Centre for Earth Evolution and Dynamics (CEED), University of Oslo, Norway.

[3]Research Centre for Arctic Petroleum Exploration (ARCEx) University of Tromsø, Norway.

The northern African margin is a seismically active region as the result of the convergence between African and Eurasian plates. The Atlasic fold belt in central Tunisia, presenting a complex structural geology, is considered as an earthquake-prone area. This work attempts to illustrate a clear case of seismic activity triggered by the Quaternary rejuvenation of a seismogenic basement fault in the Central Tunisia area. The area of interest is situated within the Altas region of Central Tunisia bordered by the North-South Axis to the east. Focal mechanisms of largest earthquakes were taken from local and international catalogs. The Kasserine fault is a NE to E-trending, seismically active right-lateral (dextral) strike-slip basement fault that extends for hundreds of Kilometers, and passes transversely through major fold structures. The inversion of focal mechanism data by using Win-Tensor demonstrates a prevailing NW-SE oriented maximum compressive stress (σ1) in the central Tunisia area. The results support the active deformation rates along the Africa-Eurasia collision zone. Earthquakes in Central Tunisia are mainly generated by the neotectonic remobilization of many seismogenic basement structures that occur as major faults or well-defined structural and topographic lineaments. Our findings are in strong agreement with the NW–SE Africa-Eurasia convergence, which is accommodated by an array of strike-slip movement along the N-S fault system in this part of Central Tunisian Atlas.

Keywords: Active faulting · Seismogenic fault · focal mechanisms · Stress inversion, Central Tunisia.

Spatial Changes and Relationship Analysis of NDVI and LST Using Satellite Image Datasets in Pottuvil Divisional Secretariat Division, Ampara

M. L. M. Hicmathulla[1], A. W. F. Nafla[2]

[1]Faculty of Graduate Studies, University of Sri Jayewardenepura, Colombo, Sri Lanka
[2]Department of Bioscience, Vavuniya Campus of the University of Jaffna, Vavuniya, Sri Lanka

Pottuvil Divisional Secretariat (DS) division has witnessed developmental activities such as building, road construction, agriculture, tourism and many other anthropogenic activities. Observable anthropogenic activities and rapid urban development cause deterioration of vegetation cover. Vegetation cover has a significant influence on the surface temperature. Identifying the changes and the relationship is a major role for the natural resource management, planning and implementation sector, and it depends on the availability of data related with the environmental parameters. The objective of this study is to map NDVI and LST using Landsat satellite image datasets using a geographic information system and to identify the changes and analyze the LST-NDVI relationship with the extracted satellite image datasets. This study was done using the satellite images to prepare Normalized Difference Vegetation Index (NDVI) and Land Surface Temperature (LST) maps, and to identify the changes and relationship by overlaying process of NDV and LST in the study area of Pottuvil DS division and done for specific years of 1991, 1996, 2001, 2006, 2009, 2014, 2018 and 2020. Results show that high temperatures were shown mainly in areas with buildings and constructions which have low NDVI values, whereas low temperatures were shown mainly in forest cover and agriculture land which has high NDVI value. R2 values were identified by the scatter plot regression analysis between the NDVI and LST as 0.38 in 1991, 0.31 in 1996, 0.21 in 2001, 0.47 in 2006, 0.28 in 2009, 0.51 in 2014, 0.50 in 2018 and 0.44 in 2020. From this R2, it could not conclude that a strong linear relationship between the LST and NDVI due to the low R2 values for all the study years, but a negative correlation was observed between LST and NDVI. This study mainly showed the ability of remote sensing and the related satellite image data to identify, monitor and manage the natural resources and environment, sustainably and effectively.

Keywords: Satellite Image Data, NDVI, LST, Spatial Change, Remote Sensing, Pottuvil DS division.

Gamified Feedback and Job Satisfaction of Generation Z Software Professional in Post-COVID-19 Environment

H D Jayathilake[1], N P Barupala[2], J A Jayathilake[3], H M Jayathilake[4]

[1]Department of Business Management, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka
[2]Department of zoology and environmental sciences, University of Colombo, Colombo, Sri Lanka
[3]Department of Computing, Informatics Institute of Technology, Colombo, Sri Lanka
[4]Division of Information Technology, Institute of Technology - University of Moratuwa, Homagama, Sri Lanka

The global expectation is that millennials and Generation Z will take the stake of 75% of the workplace population by 2025. Concurrently, Sri Lanka put its prospect on the software development organizations to be one of the major employers for the country. However, Generation Z software development employees are very much new to the industry, and very limited researches have been conducted in this area of research. Presently, under the Sri Lankan context, the sector has created 146,000 breadwinning software development professionals; and is considered the biggest employer for the highly skilled employees from the Generation Z generation cohort in Sri Lanka. However, the sector has hit with a very high employee turnover ratio due to the high global demand for workers and the hectic nature of the job. Therefore, this study identifies the conceptual framework that can be utilized to improve job satisfaction among the software professionals by having the immersion-, achievement-, social-related gamification features into the real-time feedback system of the software development organization. Moreover, based on the exploratory factor analysis and Cronbach's Alpha to identify the questionnaire with 31 items that can be used to further research on this area using a conceptual framework that was constructed through the literature review. Since visualization of the performance is very much important to demonstrate the organizations' integrity, transparency on the performance evaluation with unbiased nature to secure the positive ideology from the employees' perspectives. Gamified features are suggested as the data visualization techniques from this paper by proposing a conceptual framework with the questionnaire constructed based on the previous literature. Based on the framework, it explains the impact of feedback systems with gamified visualization to address the needs of Generation-Z to have job satisfaction while working through work from home in a post-COVID-19 environment.

Keywords: Gamification, Feedback, Job Satisfaction, Software Professionals, Generation Z, Data Visualization

A Deep Learning Approach to Hate Speech Detection in Sinhala

B. A. S. S. B. Jayawardhana, M. R. D. S. G. Punchihewa
Department of Industrial Management, University of Kelaniya, Kelaniya, Sri Lanka

Although social media platforms bring the entire world together, they can also be the cause of a numerous range of crimes. Hate speech is one of the many ways they can bring about problems within the social groups of society. It is important that these platforms are able to detect such content in advance and remove them before they have been viewed by too many users. Due to the number of languages that are supported by the current social media applications, algorithms that apply to all of them must be developed. This study aims to identify which algorithm works best on social media content in Sinhala. Here, we propose a deep learning-based approach using a model which incorporates Long Short-Term Memory (LSTM) units and FastText word embeddings. The model trained on the Sinhala Unicode Hate Speech dataset from Kaggle, which contains 6345 comments on Facebook, where 3455 (54%) of them relate to hate speech. This deep learning model has been compared with other machine learning algorithms like Naïve Bayes, Logistic Regression, Support Vector Machines, K-Neighbors and Decision Trees. It was found that the proposed model produced the best results on the data. The model uses pre-trained 100-dimensional word embeddings from the FastText library. This makes up the first layer. The layers following this include a Bi-directional LSTM layer, a Dense layer with a Rectified Linear Unit (ReLU) activation function and finally, a Sigmoid layer for binary classification. All of the models have been trained and tested on a 70-30 train-test split. By replacing the model with the appropriate embeddings, the proposed model can be re-trained and used for hate speech detection in any language. However, a potential avenue for future work would be to pursue a model for the detection of hate speech in a multilingual setting.

Keywords: Hate Speech, Deep Learning, Recurrent Neural Networks, Social Media, Natural Language Processing

Quantitative Evaluation of Equity Research for Accurate Investment Decisions in Future

D. R. S. Kalubowila[1], R. Fernando[2], S. D. Viswakula[1]

[1]Department of Statistics, University of Colombo, Colombo, Sri Lanka

[2]Specialized Solutions, Acuity Knowledge Partners, Colombo, Sri Lanka

Equity research is among the most important information sources for making investment decisions on equities. The purpose of this study is to utilize statistical measures to evaluate the estimates and recommendations forecasted in equity research reports. Measures are applied to a publicly sourced data set of 16699 forecasts by 27 brokerages covering 195 equities from 2016 to 2019 in the New York Stock Exchange & National Association of Securities Dealers Automated Quotations exchange in the United States of America spanning 5 sectors (i.e Computer and Technology (5686), Retail Wholesale (4556), Medical (2868), Consumer Discretionary (1900), Finance (1689)). Accuracy of estimates was measured with proportional mean absolute forecast error (PMAFE). PMAFE incorporates consensus allowing to make comparisons with competitors & it is adjusted for heteroscedasticity. PMAFE showed that none of the brokerages has been able to issue consistently accurate results for the period of study. Using regression for absolute percentage errors (APE) enabled measuring the consistency of estimates issued. APE time regressed slope values indicated that only one brokerage was able to issue reports that were least likely to change as the financial year progressed. The leader-follower ratio gives information on the ability of analysts/brokerages to interpret new information. None of the sampled brokerages was able to lead consistently and significantly for the period of study. Simulated portfolio formation was utilized to compare risk-adjusted portfolio returns. Out of the samples, 6 brokerages were able to generate positive relative risk-adjusted returns for the period of study. This paper has carefully selected a combination of metrics to give a 360-Degree view evaluation for equity research reports which gives an improved, unbiased and accurate evaluation compared to previous work. Thus, the framework can be utilized to benchmark analysts/brokerages by investors/institutions to make accurate and informed investment decisions in the equity market.

Keywords: Equity Research, Evaluation, Data Ethics, Econometrics, Investment

Improving Sinhala Language Modelling through Deep Learning

D. S. Kulatunga[1], A. R. Weerasinghe[2], E. R. A. D. Bandara[1]

[1]Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka

[2]University of Colombo School of Computing, Sri Lanka

Language modelling can be considered as the foundation for most upstream Natural Language Processing (NLP) tasks. Sinhala, though a language spoken by a majority of Sri Lankans, has very limited research carried out on modelling it. This research is an attempt to address that gap particularly by employing deep learning techniques. Previous studies have mostly attempted it as an initial step in addressing an upstream task such as speech recognition or Part of Speech Tagging. This study aims at using state-of-the-art deep learning algorithms in order to obtain a wide coverage language model for Sinhala. We initially used a publicly available Sinhala corpus to build an n-gram model and a Long Short-Term Memory Recurrent Neural Network (LSTM – RNN) model before turning to the state-of-the-art NLP technique, the Transformer. The Transformer used was based on the Bidirectional Encoder Representations from Transformers (BERT) model trained on 10M word corpus owing to model complexity and computing resource limitations. The n-gram and LSTM models were evaluated and compared using predictive text (human evaluation) while all models were also evaluated using the language perplexity (machine evaluation) metric. The BERT models cannot be evaluated using predictive text since transformers are masked models which predict words given their surrounding context. The human evaluation resulted in a 4-gram model trained on a 2M word corpus as the best predictive text model, while the best overall language model for upstream NLP tasks was the BERT (transformer) language model trained on a 10M word corpus resulting in a perplexity of 26.67. Training the Sinhala BERT model on larger corpus can be expected to result in significantly reduced perplexity scores, in turn improving the richness of the corresponding language model.

Keywords: Language Modelling, Natural Language Processing, Sinhala

Forecasting the Directional Movement and Value of S & P SL 20 Index Using News and Macro-Economic Indicators

M. P. Maha Arachchi[1], G. P. Lakraj[1], I. Wijesekara[2], P. Wijesiriwardana[2]

[1]Department of Statistics, University of Colombo, Colombo, Sri Lanka

[2]Acuity Knowledge Partners (Pvt) Ltd, Colombo, Sri Lanka

Stock market is one of the factors which affect the economic performance of a country by promoting investments. Stocks represent the interest of ownership in a company and trading of stocks gives the chance for companies to raise funds for their future operations. Standard & Poor's Sri Lanka 20 (S & P SL 20) index is a subset which helps investors to gain insights on the market performance of 20 leading companies listed in the Colombo Stock Exchange (CSE). This study aims to forecast the directional movement and the value of S & P SL 20 index using Natural Language Processing (NLP) and Machine Learning (ML) techniques. The directional movement has been forecasted using English news contents published on 10 selected news sites and Sri Lanka Investor Forum. The monthly closing value has been forecasted using 12 macro-economic indicators such as currency in circulation, total domestic credit, merchandise imports/exports with the sentiment scores obtained for the text data. 1st of January 2015 to 30th June 2020 has been considered as the time period for the analysis. Bag of Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF) and VADER (Valence Aware Dictionary and sEntiment Reasoner) with TextBlob have been used to analyze text data where Random Forest, XGBoost, SVM, Ridge, LASSO and Elastic-net regression were applied as ML techniques. VADER with TextBlob has outperformed BOW and TF-IDF when transforming text data into vectors. The Random Forest model has provided a better predictive ability than the other shrinkage methods in the second phase. However, the study reveals that the predictive power of news and the selected macro-economic indicators is relatively low to forecast the directional movement and the value of S & P SL 20 index.

Keywords: S & P SL 20; News; Macro-Economic Indicators; NLP; ML

Multi Target Regression for Click Through Rate Prediction

Rajitha Manellanga

Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, Sri Lanka

Digital advertising is a specific way of using digital platforms for advertising purposes. This method is popular among several companies for bringing potential customers' attention to their products and services. If we can recognize the behavior of target groups towards ad campaigns, it is beneficial to implement cost-efficient advertising strategies. Displaying ads on websites is a popular way of digital advertising. Predicting Click Through Rate, which is the ratio between the number of clicks and the number of impressions of an ad is worth, since it affects the profits of advertisers. This study proposes a multi-target regression approach to predict both the number of ad clicks and the number of impressions. Through that, corresponding click-through rates are computed. The multi-target regression approach was implemented using several machine learning techniques when target variables are highly correlated and the distribution of the ad clicks is imbalanced among several age groups. The prediction results were compared using the coefficient of determination by weighting each target equally. Since the coefficient of determination shows the proportion of the response variable's variance captured by the regression model, it was chosen over the mean squared error, which captures the residual error. The Random Forest Regressor demonstrated dominance over the AdaBoost Regressor and Decision Tree Regressor with a coefficient of determination of 0.8. Since experimental results are satisfactory, this approach is applicable to predict click-through rates in online ads. The proposed method in this study is based on several technical factors of online ads and the age of the target groups. Therefore, this approach is beneficial to build digital advertising strategies. Also, it is worth identifying the patterns of both the number of impressions and the number of ad clicks rather than predicting the click-through rates directly.

Keywords: multi target regression, machine learning, click through rate, advertising

Investigating the Chest Imaging Features of Covid-19 Patients using Deep Neural Networks

A. B. Meepaganithage, M. G. N. A. S. Fernando

University of Colombo School of Computing, University of Colombo, Colombo, Sri Lanka

The COVID-19 pandemic causes devastating effects on global health, the economy, and the livelihood of people. The timely diagnosis of COVID-19 is crucial to control the spread. The need for additional testing methods has increased due to the limitations in current testing methods. Some of the studies showed that medical imaging techniques can be used to detect COVID-19. In this study, we used deep learning methods to classify chest X-rays as COVID-19, normal, and pneumonia. We developed two deep learning models to detect COVID-19 using Posteroanterior (PA) view and Anteroposterior (AP) view chest X-rays. Two datasets of 300 chest X-rays (100 healthy, 100 Pneumonia, 100 COVID-19) for PA view and AP view were used. As the first deep learning model, a new Convolutional Neural Network was built from scratch. Then, VGG16, VGG19, and ResNet50 transfer learning models were used. Finally, the transfer learning models were extended by adding more layers to the top of the existing model. As the first part of this study, we used PA view chest X-rays and obtained 98% overall accuracy, and 98% precision, 99% recall, and 99% f1-score for the COVID-19 class using the extended VGG19 model. In the second part, we used AP view chest X-rays and obtained 79% overall accuracy, and 96% precision, 83% recall, and 89% f1-score for the COVID-19 class using the extended ResNet50 model. Finally, gradient-based class activation maps were generated using the extended VGG19 model to visualize the areas that helped the model in detecting COVID-19. The average time required for training and testing were 31 minutes and 2 minutes respectively. This research showed that higher performance can be obtained in detecting COVID-19 using extended transfer learning models. In PA view chest X-rays extended VGG-19 model and in AP view chest X-rays extended ResNet50 model performed the best.

Keywords: COVID-19, Chest X-ray, Machine Learning, Deep Learning, CNN, Transfer Learning

Topic modeling in Youtube Data

A. C. Nanayakkara, G. A. D. M. Tennakoon

Sabaragamuwa University of Sri Lanka

The topic comprehends a cluster of words that often happens together. Topic modeling is a method for unsupervised classification of documents, similar to clustering on numeric data, which finds some natural groups of items (topics). It gives a prominent way to analyze and categorize big unclassified text. The main objective of the research study is to consider a topic modeling problem, and examine two popular topic modeling methods in python i.e. Latent Dirichlet Allocation (LDA) and K-Means Clustering, utilizing tools in the scikit-learn and gensim packages. The study shows how these methods can be utilized to perform topic modeling using the same data set, together with common preprocessing steps in the analysis. The data was collected from YouTube videos by using the latest YouTube Data Application Programming Interface (API) V3. Corpus preparation and cleaning were performed using Python Natural Language Toolkit (NLTK) that provides stop-word removal, stemming, lemmatizing, tokenization, identifying n-gram procedures, and other data cleanings like lowercase transformation and punctuation removal. After the preprocessing step, CountVectorizer followed by the TfidfTransformer has been utilized to process the data. The study discusses the advantages and drawbacks of each method. K-means clustering using vectorized data is one method for clustering similar documents using the distance between the document vector representations. Latent Dirichlet allocation is a generative model, where documents are viewed as a mixture of topics, and topics as a mixture of words. The study has revealed that determining the "right" number of topics can be a challenging task, yet it is possible to utilize measures such as the silhouette score for k-means clustering or coherence for latent Dirichlet allocation and by analyzing the overlap of words in each topic can help to determine what a good value for the number of topics is.

Keywords: Topic modelling, YouTube comments, Latent Dirichlet Allocation (LDA), K-Means Clustering

Predictive Analytics for Machinery Maintenance in the Production of Quality Garments

U.M.M.P.K.Nawarathne, Vijani Piyawardana, Jesuthsan Alosius

Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

The garment industry in Sri Lanka has brought an immense economical advantage to the country in recent years. With its growing popularity, this industry has faced a lot of issues that could affect its applicability in earning profit. Proper functioning of machinery is one such crucial function that has to be considered in the factories when producing quality garments effectively. Two main issues found in machinery operations are unexpected machinery failures and lack of knowledge regarding cycles run by machines in the future. Observing these facts, it is clear that predicting a problem before it occurs and understanding their future behaviour beforehand are both wise decisions. Hence the main objective of this work was to provide the machinery crews with a mechanism to prepare for the sudden problem occurrences using early predicted information. Statistics regarding machinery information, non-breaking errors thrown by them, component failures, maintenance records, and machinery sensor readings were taken into account as data to study this use case. Hence this research paper presents a methodology that has used Predictive Maintenance with Multi-Class Classification and Random Forest Classifier to predict machinery failures and KNeighboursRegressor classifier from Regression Analysis to compute the remaining lifecycles run by machinery thereby providing statistics and future outcomes in advance. After a series of testing the finalized models, Random Forest Classifier gave an accuracy score around 97.8% while the KNeighboursRegressor gave a variance score around 0.88. The outcome of this research gave early machinery component failure predictions and their future life cycle counts as the results which were used to build a software solution that is comprised of features such as having the ability to visualize the above-mentioned predicted insights along with the other necessary graphs and dashboards to view the corresponding statuses of machinery which will be assisting the needs of machinery crews.

Keywords: Machine Learning, Predictive Maintenance, Multi-Class Classification, Regression Analysis

Study on Part-Time Employment of Science Undergraduates, University of Colombo

K. G. J. Nishadini, R. A. B. Abeygunawardana

Department of Statistics, University of Colombo, Colombo, Sri Lanka

This study aims to identify the factors associated with students' part-time employment and to identify students' perceptions about part-time employment. A cross-sectional online survey was conducted to gather the relevant information through a questionnaire. All the students in the Faculty of Science, University of Colombo who choose to engage with part-time jobs or not, were considered to capture the aspect of interest. Stratified random sampling was used by considering four academic years as strata. The objectives of the study have been achieved by using the XGBoost machine learning technique which is a tree and rule based ensemble method that can be considered as an intrinsic feature extraction method. It was found that, based on the model accuracy and precision, the XGBoost was the best algorithm. Engaging with extracurricular activities, type of financing, place of residence, gender, father's higher education level, first priority on expenses, number of dependents in the family, stream, age, and time spent studying outside of university were identified as the most important factors associated with part-time employment. The model accuracy was 86.54% with 66.67% precision. Further, Students' perception about the part-time employment was considered by using a likert scale from strongly disagree to strongly agree. The students agreed that the university should be involved with part-time jobs and that it would be more effective to engage in part-time jobs relevant to the field of study. However, the employed students have no clear idea that their part-time job will help them to clarify their future career. They also claim that their work cannot have a positive impact on their academic performance, but can have a positive impact on their social lives.

Keywords: Science Undergraduates, Part-time Employment, Machine Learning

Machine Learning Approaches to Forecast Inflation in Sri Lanka

W. P. I. S. Pathirana, C. D. Thilakaratne

Department of Statistics, University of Colombo, Colombo, Sri Lanka

Inflation is the measure of rise in the prices of most goods and services of daily or common use such as food, clothing, housing, transport etc. The practice of forecasting inflation is generally being considered as an important task since it is very beneficial for the economy of a country. It will be really important in the decision making phase for a developing country like Sri Lanka. Moreover low inflation is a benefit for the natives and for the investors, because Sri Lankan economy is a free market economy, where investors locally and globally are interested. This study aims to forecast inflation rate on short term horizons in Sri Lanka using various types of machine learning techniques. The existing literature in Sri Lanka has investigated forecasting inflation rate using benchmark models. The use of machine learning techniques to forecast inflation in Sri Lanka is very rare. Hence, this study contributes with new research findings by exploring various types of machine learning approaches to forecast Sri Lankan inflation rates. CCPI is used as the measure of inflation rate and GDP, merchandise imports and exports, monetary aggregates (M1 & M2), Tea production, Rubber production, Coconut production, Marine Fish production, Total tourist arrivals are the predictor variables which were chosen after referring to past literature. Ridge Regression, Lasso Regression, Elastic-Net Regression, Random Forest(RF), XGBoost, Support Vector Regression(SVR) , K-Nearest Neighbour(KNN),Long Short Term Memory (LSTM) were applied in the study as the machine learning techniques. RMSE, MSE, MAE and MDA were used as the accuracy measures for the applied machine learning models.

Keywords: Machine Learning; Random Forest; XGBoost; Long Short Term Memory; Colombo Consumer Price Index (CCPI); Inflation rate

A Comparative Study on Artificial Neural Network Techniques for Short-term Rainfall Forecasting in Colombo, Sri Lanka

G. C. J. Piyatilake, K. A. D. Deshani

Department of Statistics, University of Colombo, Colombo, Sri Lanka

Rainfall forecasting remains a challenging topic for many decades because of its dynamic, complex, and nonlinear nature. In recent years, the application of Artificial Neural Networks(ANN) on climate variables has been popular among researchers because of their ability to handle more volatile, nonlinear behaviour of climate-related data without considering any prior assumptions. The objectives of this study are to compare the performances of different ANN techniques for short-term rainfall forecasting in Colombo and to experiment on the model performances considering the number of steps ahead of the forecast and the length of the training period. Four different ANN architectures; Feed-Forward NN, Long Short-Term Memory(LSTM), Bi-directional LSTM, Gated-Recurrent-Unit(GRU), have been implemented to obtain one-day-ahead forecasts and compared their performances. Daily data for rainfall, atmospheric pressure, wind direction, wind speed, cloud cover, minimum and maximum temperature in Colombo for 10 years were used in the study. Additionally, variables monsoon-season and month-of-the-year were also considered in the analysis while wind direction and wind speed were combined to get the effect of the wind in a more meaningful way. To identify the most promising features that are highly associated with rainfall, two feature-selection techniques, namely Mutual Information and Random Forest(RF) have been used. Among the four models, the LSTM model trained with selected features from RF was the best-performed model for this dataset with the lowest RMSE of 14.203mm. Then the best model was used to study the model performances by changing the number of years considered for model training. It was observed that overall, RMSEs decreased as the time length increased and the lowest RMSE was recorded when using seven years-long training data. Finally, the best model was further applied to find the number of days the model can predict efficiently into the future and found the maximum number of four days can be forecasted without changing the model performances drastically.

Keywords: Rainfall forecasting, LSTM, feature selection, multi-step forecasting

Game Outcome Predictor and Game Plan Generator for Twenty20 Cricket

L. C. P. Pussella[1], R. M. Silva[2], W. C. P. Egodawatta[3]

[1]Department of Physical Sciences, Rajarata University of Sri Lanka, Mihintale, Sri Lanka

[2]Department of Statistics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka

[3]Department of Plant Sciences, Rajarata University of Sri Lanka, Anuradhapura, Sri Lanka

As its main aim, this research study attempts to develop an in-game outcome predictor by targeting the Twenty20 (T20) cricket games played in the Indian Premier League (IPL) from 2008 to 2014. Since T20 cricket is a format of the game in which the final outcome is decided by a number of factors, 40 candidate features were initially recognized as the elements having an intuitive significance towards the final result of a game. Afterwards, by considering the Least Absolute Shrinkage and Selection Operator (LASSO) as a feature selection method, the candidate features were filtered to obtain a subset of features

having the highest levels of importance. The filtered features were utilized as the next step for creating three distinct classification models based on Naïve Bayes, Logistic Regression and Support Vector Machines (SVM) and each classification model predicts the game outcome at the conclusion of the first half of any given IPL game. Throughout the study, prediction accuracy is used as the performance metric for evaluating the performance of each classification model where Naïve Bayes classifier performed with an accuracy of 72.58%. In contrast, both Logistic Regression and SVM demonstrated an accuracy of 67.74%. Furthermore, this study aims at introducing a Game Plan Generator of which, the objective is to provide the support to the playing teams in the second half of the game. It guides the teams by offering a set of specific combinations of values for features corresponding to the second half of the game, along with the respective winning probabilities of such combinations (game plans) for the team batting second. Finally, with the aim of enhancing the visibility and usability of the study, we have developed a web-based application which presents the overall outcome of the study via an interactive dashboard.

Keywords: Cricket, Sports Analytics, Machine Learning, Naïve Bayes, Logistic Regression, Support Vector Machines

Machine Learning Approach to Detect Tea Plant Diseases in Sri Lanka

S. N. K. Rajakaruna[1], Rangana Jayashanka[2], H. A. S. G. Dharmarathne[1]

[1]Department of Statistics, Faculty of Science, University of Colombo, Colombo, Sri Lanka

[2]University of Colombo School of Computing, Colombo, Sri Lanka

Agriculture is the primary source that supplies food for every human being. Apart from being a food supplier, agriculture provides many benefits. For most countries in the world, agriculture plays a vital role as a foreign exchange resource. Agriculture also plays a leading role in Sri Lanka's economic sector. Tea cultivation is a primary agricultural activity in Sri Lanka that adds foreign currency to the economy. Currently, tea leaf diseases are one of the major problems faced by the Sri Lankan tea industry. Therefore, one of the compelling requirements of this field is the early and reliable identification of tea leaf diseases. Thus, this study was carried out to provide a solution to this problem by providing a machine learning model to accurately identify the tea leaf diseases. This study began by collecting the images of tea leaves with diseases and without diseases from several tea estates in Sri Lanka. All the collected disease leaves are of disease blister blight. After collecting the images, those images were undergone through several pre-processing steps. After pre-processing the images, multiple machine learning models were fitted to classify the tea leaves. Logistic Regression and Random Forest models outperformed the other machine learning algorithms by giving a 56.15% accuracy. Then from the K-Nearest Neighbor (KNN), 53.85% of accuracy was obtained. 50% of accuracy was obtained from the Support Vector Machine (SVM), and the accuracy obtained from the Decision Tree model was 48%. Convolutional Neural Networks (CNN) were fitted by changing the number of convolutional layers to increase the model accuracy. With five convolutional layers, CNN accurately classified the tea leaves as disease and non-disease with an accuracy of 97.69%. Therefore, the implemented CNN will support the tea industry of Sri Lanka to identify the tea leaf diseases accurately.

Keywords: Agriculture, Sri Lanka, Tea leaf diseases, Machine Learning, CNN, Convolutional layers

Question Answering For Sinhala Language Using Deep NN Architectures

M. W. K. S. Rasantha[1], Dr. A. R. Weerasinghe[2], Dr. J. H. D. S. P. Tissera[1]

[1]Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka

[2]School of Computing, University of Colombo, Sri Lanka

Question Answering (QA) is an increasingly researched area in Natural Language Processing and has gathered a lot of attention in recent years due to its commercial potential. However, there is a lack of resources in many languages including Sinhala for carrying out Natural Language Processing research. In particular, the amount of research done for Sinhala QA is minimal. This study was carried out to address this gap and propose a method for modelling Sinhala QA using Transfer learning techniques. A dataset was created to facilitate Sinhala QA by translating part of the SQuAD 1.0 English QA dataset using the Google translation API. The translated dataset was post processed to reduce translation errors. A considerable amount of translated text was also omitted due to the "Span mismatch problem". The final dataset stands as the first Sinhala reading comprehension dataset. The proposed model is a Multilingual T5 model, fine-tuned for Sinhala QA. It resulted in an F1 score of 54.96%. The results were promising owing to this being the first QA model for Sinhala and sets a benchmark for developing state of the art Sinhala QA models in the future. Hyper-parameters had to be chosen according to the hardware constraints rather than any convergence criterion. Evaluation of multilingual models for Sinhala QA was also carried out. XML-Roberta and Multilingual XML models had the highest F1 scores of 62.47% and 51.26% respectively among the fitted multilingual models. By removing the computational infrastructure limitations, the performance of the system is expected to be significantly increased.

Keywords: Natural language processing, Question Answering, Sinhala language, Computational linguistics, deep neural network architectures, Transfer learning

Algorithm to Detect Doctored Images

R. M. D. S. Rathnayaka[1], S. D. Viswakula[1], T. A. M. C. Thantriwatte[2]

[1]Department of Statistics, Faculty of Science, University of Colombo, Colombo, Sri Lanka

[2]Specialized Solutions, Acuity Knowledge Partners Pvt Ltd, Colombo, Sri Lanka

Fake images and fake news have been a disaster for the entire world in the past few decades. One of the biggest problems was the socialization of fake images. Doctored image classification is a vital process of controlling the increment in image forgeries. In a world of highly advanced technology, automation of doctored image classification using image processing, machine learning or deep learning has been a novel approach and interesting research area. Doctored images differ in their background, objects, and doctoring methods in the images. Visual classification of the doctored image with some pixel-based, camera-based, and physics-based methods underperformed due to the high variation in doctored images. To address this problem, an algorithm is proposed to reduce the accuracy drop in previous methods. The CASIA image database which is often used for research was used for this experiment. From CASIA 2 image dataset 5136 Authentic images and 5136 Doctored images were used. Under the feature extraction, all images were converted into *Numpy* arrays in python. 97.6% accuracy was obtained in a previous study using CASIA 2 dataset. The traditional machine learning models were applied first. Under the machine learning models Logistic regression and Decision Tree gave higher accuracy than Random Forest, Gradient boost, and Support Vector Machine model with and without hyperparameter tuning. Decision Tree and Logistic Regression models were given 50% of accuracy with 50% precision and 50% recall and there was a big accuracy drop in traditional machine learning models. For further improvements, Deep Learning and Transfer Learning models were added. VGG16 model and AlexNet models were given slight improvement for doctored image classification. The alexNet model was able to gain an accuracy of 79%. Convolutional Neural Network was applied finally to the model. To get better results several Convolutional layers were added and trained the model, through which the accuracy was improved to 92%.

Keywords: Digital Image Forgery, Copy-Move Forgery, Image Splicing, Image Retouching, Machine Learning, CNN

Developing the Modelling Framework of an Intelligent News Recommendation Engine for Financial Analysts

H. M. P. P. K. H. Samarasekara[1], W. M. N. D. Basnayake[1], D. Hureekaduwa[2], B Weralupitiya[2]

[1]Department of Statistics, University of Colombo, Colombo, Sri Lanka
[2]Acuity Knowledge Partners, Colombo, Sri Lanka

Referring to news articles has become a daily routine of each financial analyst, as it helps to make accurate financial insights. However, as there are thousands of news articles generated daily, finding out the most relevant news articles for the financial analyst becomes a time consuming task. Hence this study develops a modelling framework of an intelligent news recommendation engine for financial analysts which assists to recommend the most appropriate articles according to analysts' preferences in an efficient and effective manner without tedious browsing. The data collection phase of the recommendation engine is accomplished through retrieving news articles from online news websites using web scraping technique. As the initial step, the response variable, analysts' preference level for each article is obtained from a group of financial analysts at a selected financial company. As a solution to the imbalance problem exists in the obtained dataset, the text data augmentation method which synthetically generates new copies of data based on the already available data is applied. The analysis phase is carried out using the NLP approach. NLP preprocessing techniques, punctuation and noise data removal, lowercasing, stop word removal and lemmatization are employed as refinements to the data sparsity problem. Three machine learning (ML) techniques, KNN, SVM, Random forest with BOW and TFIDF feature extraction methods and two deep learning (DL) models LSTM and CNN are utilized on the obtained dataset. The DL model, CNN obtains the highest accuracy with comparison to other applied methods. Hence the CNN algorithm is chosen as the most suitable approach in the analysis phase of the recommendation engine to filter out the relevant news articles in the selected financial company. Moreover, this study reveals that DL models perform considerably higher performances in the context of news recommendation engines rather than ML approaches.

Keywords: Financial Analysts, Recommendation engine, Natural Language Processing, Data Augmentation, Machine Learning

Improving Bank-Customer Efficiency through Automatic Speech Recognition

L. D. P. S. Senaratne[1], H. N. D. Thilini[2], C. H. Magalla[1]
[1]Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka
[2]University of Colombo School of Computing, Sri Lanka

Automatic Speech Recognition (ASR) and deep learning have become state-of-the-art technologies in today's world. From the technological sector to the finance sector, speech recognition plays an important role. On top of that, industries are looking into technology-enabled automation of complex business processes; specifically Robotic Process Automation (RPA). In particular, if speech recognition was to be implemented in the banking industry, it would improve the efficiency of both the bank and its customers. Therefore, automation of the filling of bank-slips has been proposed. The focus is on the number of entries done on the bank-slips because they are a main part of the day-to-day transactions that happen in a bank. The numbers utilized include bank account numbers, credit card numbers, telephone numbers, national identity card numbers and amounts in Sri Lankan rupees. This research is aimed at building an ASR system to identify such numbers used in banks. The system is designed to identify numbers of any given format in Sinhala language, using the Kaldi toolkit. The dataset is a primary speech dataset consisting of audio files of 51 females and 49 males. The experiments are conducted using a phonetically balanced audio dataset that contains approximately 6 hours of speech data. The test set includes 1 hour of audio data of 11 females and 9 males. First, a statistical model of Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) is trained as a base-line model. Then, the experiments are carried out on Deep Neural Network-Hidden Markov Model (DNN-HMM) using tanh non-linearity and p-norm non-linearity. The performances of the Deep Neural Network (DNN) models are compared with the statistical baseline model at the end. The results indicate that the DNN architecture with tanh non-linearity performs the best with a Word Error Rate (WER) of 4.81%. Hence, the introduced system could be implemented as a banking application.

Keywords: Banking sector, Sinhala numbers, ASR, DNN, Kaldi, RPA

Vehicle Detection In Satellite Images

K. Sutharsan[1], S. Gamika[2], Dr. G. P. Lakraj[1]

[1]Department of Statistics, University of Colombo, Colombo, Sri Lanka

[2]Acuity Knowledge Partners, Colombo, Sri Lanka

In recent years, vehicle detection in satellite images has been a big challenge and an active field of study. Identification of small objects such as vehicles in satellite images is a very difficult task due to the complex background, varying colours and occlusions caused by buildings and trees. The focus of the research is to detect vehicles in satellite images in real time and to study the effect of resolution in detecting vehicles. You Only Look Twice (YOLT) model was chosen for this study and the Cars Overhead With Context (COWC) data set was used which is a vast collection of annotated cars from overhead. Two experiments were carried out by changing the number of training images and iterations to assess the performance of the model . Experiment 1 used all the 1683 image cut-outs for 2000 iterations and experiment - 2 used 250 image cut-outs for 4000 iterations. Experiment - 1 performed better with a Mean Average Precision (mAP) value of 0.59. The results showed that increasing the iterations without a fair amount of training data will not increase the performance of the model. Due to the limitations of the resources, resolution study was conducted with You Only Look Once (YOLOv5) model. Two datasets were prepared with 1024 x 1024 and 2016 x 2016 pixels size. Further, gaussian kernel with three levels of sigma values were used to blur the images. The evaluation results showed a gradual decrease in the mAP with increased blurring. Greater mAP was achieved with 2016 x 2016 size images without blurring, indicating that higher resolution images perform better. For the future studies, we suggest conducting a systematic approach instead of the ad hoc approach in the experiments for better understanding. Further, we suggest conducting the resolution study with YOLT model by utilizing the resources properly.

Keywords: vehicle detection; satellite images; computer vision; deep learning; investment research

Linkage Comparison on Agglomerative Clustering for Online Retail Dataset

R. M. B. P. M. Uduweriya, N. A. D. N. Napagoda

Department of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka

Online retailing is the sale of goods and services over the internet which is growing at an astonishing rate, with online sales currently accounting for the total retail market which leads to gathering an excessive amount of retailing data. With the rapid production of data in many forms, online retail data cannot be managed deftly, which ended with copious raw data. Therefore, grouping the online retail data into a set of similar homogeneous groups is extremely significant. However, unfortunately, false grouping occurred with noisy data. Therefore, this study proposes a way to cluster the online retail data set with a comparison of linkage methods to make the clustering process more effective. For the experiment of this research, a standard data set of online retails stores are used. This data set consists of five lakh records of retail transactions with eight attributes which are invoice-no, stock-code, description, customer, quantity, unit-price and country. For improvement in the quality, data prepossessing is performed before clustering. Gower algorithm is applied to preprocessed data since it supports both text and numerical data. A Random sampling method is used to obtain samples that suit the maximum storage capacity of the computer to generate dissimilarity matrices due to the lack of storage facility to store the dissimilarity matrix. After that, the Gower algorithm is implemented to the best sample through 10 random samples which consume less run time to find the dissimilarity matrix. After that, the agglomerative coefficients of four major hierarchical clustering linkage methods are compared. The Elbow method is operated to determine the optimal number of clusters. The Elbow method was performed for each 8000 random sample dataset and the result was the same. Finally, the Ward method is the best linkage since the agglomerative coefficient was maximum in the Ward linkage.

Keywords: Dissimilarity matrix, Gower algorithm, Clustering, Elbow method, Linkage

Building an OCR Tool with a Better Text Recognition for Financial Domain

B. L. A. S. Upekshika[1], S. D. Viswakula[1], M. F. S. Ahamed[2], F. L. Muthukumarasamy[2]

[1]Department of Statistics, University of Colombo, Colombo, Sri Lanka
[2]Specialized Solutions, Acuity knowledge Partners, Colombo, Sri Lanka

Optical Character Recognition (OCR) technology which is stated as the core technology for automatic text recognition has become a popular topic with the evolution of the world. OCR technology has created a chance to work with natural language data in documents. Therefore, Natural Language Processing (NLP) was used to understand and handle this computer encoded natural language data. This research has been mainly focused on building an OCR tool with a better text recognition for financial data. The study has used different types of image pre-processing techniques and NLP techniques to enhance the Tesseract text recognition. Tesseract is an open-source OCR engine which was trained on a wide variety of languages to extract printed text from images. The NLP techniques used in this study include Bidirectional Encoder Representation from Transformers (BERT) model, algorithm with SpellChecker module, and Hidden Markov Model. Evaluation of the OCR system, BERT model and SpellChecker module algorithms have been used by calculating cosine similarity between algorithm outputs and manually type data. Hidden Markov Model evaluation is done by using a simple proportion method on predicted data and manually type data. Fifty input datasets related to the financial domain were used as the primary input. The OCR system which was built using Tesseract and OpenCV, was able to achieve better performance in financial text recognition and its outputs have given an average of around 0.96 cosine similarity. This study has used OCR system outputs on NLP techniques separately, to enhance the text recognition by correcting misspellings. Through the evaluation, it was possible to show BERT combined with dictionary lookup method has a better performance in text correction for inputs which has less than 512 strings than the SpellCheker module. In future, solving limitation problems that occurred in the Hidden Markov Model such as improving the number of inputs will give a positive aspect on spelling correction for the financial domain.

Keywords: Natural Language Processing for misspelling correction, Tesseract text recognition, Optical Character Recognition for financial documents

Spotting Railway Signs to Build Smart Decision Support Tools in Railway Management Systems

P. T. Weerasinghe, M. Shaheer, R. Rumalshan, P. Gunathilake, E. Dayarathne

Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, Sri Lanka

Railway being an important mode of transportation, which demands highly precise management and decision support as it is extensively used for both commuter and cargo transportation. It is also considered as a salient element in smart city and modern infrastructure planning. Railway track diagrams are available with service providers in portable document format (PDF) where a single document consists of information from one station to another, including information regarding the tracks, signals, crossovers, switches, and their location details denoted, using a standard set of symbols that are drawn using a computer- aided tool. A Management tool that has all details of individual symbols is an important tool for decision support systems. This research focuses on developing an automated system to extract this information based on deep learning techniques. Here the dataset was extracted from a single PDF obtained from a leading service provider which contained fifty six images and used for training and validating the model. The method consists of two steps: object detection and optical character recognition (OCR). State of the art Convolutional Neural Network (CNN) architectures are used to perform object detection. They include single-stage detectors like YOLOv3 and SSD and two-stage detectors like RFCN and Faster-RCNN. RFCN resulted in the highest accuracy with the minimum loss value of 0.22, compared to other methods. RFCNs architecture caters to small object detection by dividing the image into small feature maps. Then, OCR is performed on detected Regions of Interest (RoI) to extract and store the text in a dedicated database which has the information of all the signs along with their location details. Image processing techniques such as template matching and Neural Network (NN) based OCR were used. Out of these two approaches, NN based technique outperformed template matching drastically with more than 50% accuracy.

Keywords: object detection, railway sign detection, railway management systems, decision support tools

A Case Study of Creditworthiness Prediction at the Loan Approval

B. N. Weralupitiya and R. V. Jayatillake

Department of Statistics, University of Colombo, Sri Lanka

Creditworthiness is the lender's willingness to trust his borrower to pay debts. This case study was carried out for a specific bank in Sri Lanka with the objectives of developing a predictive model to assess the creditworthiness of potential loan applicants and to construct an index to provide a quantitative value for each customer's credit risk. The data used for this study consisted of bank loan details of 10,626 existing customers in their loan portfolio and their repayment behaviour over 2.5 years. It consisted of 18 variables including customers' demographic and personal financial details and bank-specific ratios. Furthermore, it included 10 transaction variables that represent the quarterly default status of borrowers. The "Loan Status" was used as a dichotomous response variable with categories as "Performing" and "Non-Performing". The univariate tests such as Mann Whitney and Chi-Square Independence Tests and graphical analysis identified that apart from Customer's age and CRIB status at approval, all other variables showed a significant relationship with the response variable. As only 33% of respondents were non-performers, the Synthetic Minority Oversampling Technique (SMOTE) was used to handle the class imbalance as this method does not lead to any data loss. Machine learning techniques such as Logistic Regression, Random Forest, Support Vector Machine, and Artificial Neural Network (ANN) were applied with and without SMOTE Sampling to achieve the optimal model by comparing ROC-AUC values. ANN model applied with SMOTE sampling was found to be the best model with a ROC-AUC value of 91.6%. Further, a credit index was developed with Factor Analysis applying the Principal Component Method with transaction data. The constructed index ranges from 0-100 and it indicates, higher the index higher the risk of default. These research outcomes can be applied for any bank with minor changes and overcoming limitations like data period and data frequency issues.

Keywords: Creditworthiness, Machine Learning, Non-Performing Loans, Synthetic Minority Over Sampling Technique, Factor Analysis.

Quantifying the Heterogeneous Close Contact Patterns in Sri Lanka

W. A. P. S. Wickramarachchi, L. Munasinghe

Software Engineering Teaching Unit, Faculty of Science, University of Kelaniya, Sri Lanka

Social contact patterns are a critical factor for the spread of infectious diseases such as COVID-19. Recent research works have used different methods such as contact surveys, sensors and IoT technologies to gather social contact data. In the present research, an online survey is used to collect social contact data (from 2000 participants) to estimate age-stratified social contact matrix for Sri lanka. This survey was carried out in the Western Province (WP) of Sri Lanka because the highest number of COVID-19 cases are being reported in WP. According to the survey responses, contacts are grouped into 4 discrete age categories (0–12, 12–30, 30-60, and 60+ years). Age groups were selected according to the age stratification used in COVID-19 vaccination program in Sri Lanka. The collected data is used to estimate the so-called social contact matrix $\{m_{ij}\}$ which represents the mean contact rate between an individual in age group j with individuals in the age group i on a given day. Maximum likelihood method was used to estimate $\{m_{ij}\}$. To quantify $\{m_{ij}\}$, we adopted an assumption of reciprocity, i.e., the mean number of contacts that an individual in age group i has with individuals in age group j is equal to the number that an individual in age group j experiences with individuals in age group i. Thus the asymmetry of the contact rate was adjusted using the age-dependent population size of Sri lanka. Social contact matrix is validated using the age-stratified final sizes of COVID-19 cases at the end of its first wave. The experimental analysis shows that social contact data are useful for parameterizing the heterogeneous transmission models of various infectious diseases. Further, age-dependent assortativity, especially on weekdays implies the potential effectiveness of lockdowns or limiting the mobilization of people to mitigate the outbreaks such as COVID-19.

Keywords: Contact Matrix, Maximum Likelihood Method, Contact Survey

Workflow Composition with Predictive Modelling

C. R. Wijesinghe[1], A. R. Weerasinghe[2]

[1]University of Colombo School of Computing, University of Colombo, Colombo, Sri Lanka

[1]Faculty of Graduate Studies, University of Colombo, Colombo, Sri Lanka

[2]University of Colombo School of Computing, University of Colombo, Colombo, Sri Lanka

In workflow composition, different software components are weaved together to express a computational experiment in a workflow. Software components are heterogeneous and complex with many constraints on their parameters. Visual workflow systems such as Galaxy automate the workflow composition, enabling users to select components from a library of tools and link them in an interactive workflow development environment. Even if visual workflow systems are used, selecting appropriate tools among a long list of tools is still challenging. As a result, choosing appropriate, compatible, state-of-the-art components in workflow composition is not a trivial task. Composing workflows is a time-consuming and cumbersome task for 'bench scientists'. The objective of this study is to develop a suggestive system that will predict the next tool in composing end-to-end workflows consisting of complex scientific components. An N-gram model using maximum likelihood probabilities was developed using the workflow histories to predict the next tool in workflow composition. Around 900 workflows were collected from myExperiment workflow repository and public Galaxy servers. NLTK library was used in developing the model. Category-based prediction was done as a solution to data sparsity. Add-k-smoothing was used to handle the events with zero counts. Parameters of the model trained using a training dataset containing 80% of the tokenized data and the remaining 20% unseen dataset used in evaluating the model with 5-fold cross validation as a resampling method. The evaluation metric used is average log-likelihood. Prediction accuracy can be improved by adding more context to the prediction model. An n-gram based suggestive system built using best practice workflow histories can assist domain users in easy workflow composition. Such a system will enable domain users to develop complex state-of-the-art workflows with higher quality.

Keywords: Predictive modelling, N-grams, Workflow composition, Galaxy workflow system, Bioinformatics

Developing a Model to Decrease Call Volume Cost of a Telecommunication Company by Analysing Time-to-Failure since the Installment of the SetupBox

K. A. C. N. Wijewardana[1], H. A. S. G. Dharmarathne[1], P. A. D. L. Samarakoon[2]

[1]Department of Statistics, University of Colombo, Colombo, Sri Lanka

[2]Business Intelligence & Analytics, Dialog Axiata PLC, Colombo, Sri Lanka

The call volume cost of a leading telecommunications company is increasing day by day due to consumer complaints. Setup Box (STB) failures are highly impacted on the call volume. Thus, predicting the time-to-failure since the installment of the STB can mitigate such a concern since the company can identify how long one STB can be adopted without any failure since the installment. Moreover, essential solutions or require arrangements to abbreviate the STB failures and complaints can be fetched. Thereby the company can shrink the call volume cost. Under the main objective of the study, a model was developed to identify the time-to-failure since the installment of the STB. The response variable of interest has been the time-to-failure since the installment of the STB with the following three levels; below 10 months, below 1.5 years, and above 1.5 years. Hence, this was executed as a classification problem. Few popular supervised learning models including multinomial logistic regression, KNN, Random Forest, Decision Tree, XGBoost and Gradient boosting have been applied on the data. In order to identify the relationship between the response variable and other predictive variables, exploratory data analysis and chi-square test were performed. Further, hyper-parameter tuning was conducted to achieve the optimized parameters. In addition, several sampling techniques such as oversampling and undersampling were used to overcome the limitations caused by the class imbalance phenomena. The ratio between the train and test data was 80: 20. The indication of results show that the XGBoost has the highest classification performance, and have achieved the best predictive models under SMOTE technique based on the ROC curve interpretation.

Keywords: Machine learning, supervised learning, Time to failure prediction

Center for Data Science
University of Colombo, Sri Lanka

Center for Data Science (CDS)

Data Science is an emerging field that has capacity to grow and provide many opportunities for research and collaborative projects. Therefore, in 2016, **Center for Data Science** was established under the Department of Statistics, Faculty of Science, University of Colombo. The Center strives to facilitate research and development in Data Science in Sri Lanka through collaborations with local and international expertise both from academia and industry. It also conducts training programs, workshops and public talks to disseminate knowledge in Data Science and increase awareness of Data Science among the community. The Center promotes partnerships with local industry through consultancy projects providing them with technical expertise while enhancing skills of the students and academics in the application of Data Science techniques in the real world.