Department of Statistics,
University of Colombo

Center for Data Science

ICDS 2023

PROCEEDINGS OF THE INTERNATIONAL CONFERENCE IN DATA SCIENCE 2023

Data Science in the Age of Artificial Intelligence

19th - 20th December 2023
Department of Statistics, University of Colombo
Colombo, Sri Lanka

Proceedings of the International Conference in Data Science 2023 (ICDS 2023)

19th - 20th December 2023

at the

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

ICDS 2023 is organised by the Center for Data Science jointly with the Department of Statistics, University of Colombo

Publisher: Center for Data Science, University of Colombo

ISSN: 3030-7163

Proceedings of the International Conference in Data Science 2023
19th - 20th December 2023
Department of Statistics, University of Colombo,
Sri Lanka

Date published: 20th December 2023

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

ISSN: 3030-7163

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 Co-Chairs

It is with great enthusiasm that we warmly welcome you to the International Conference in Data Science 2023 (ICDS 2023) organised in Sri Lanka for the second time by the Center for Data Science in collaboration with the Department of Statistics, Faculty of Science, University of Colombo. We are honored to present this significant event, scheduled on the 19th and 20th of December 2023, under the theme "Data Science in the Age of Artificial Intelligence."

Reflecting the transformative connection between Data Science and AI, this conference serves as a catalyst for innovation, industry transformation, and the evolution of intelligent systems. We are excited to bring together the industry professionals, academics, and students in a hybrid format, with two pre-conference workshops, mini-hackathon and the inauguration ceremony at the Faculty of Science, University of Colombo, followed by online technical sessions.

The programme features engaging sessions, including four keynote speeches and five technical sessions covering a spectrum from Data Science and AI.

We earnestly wish that your time at this remarkable conference proves to be not only educational and productive but also genuinely enjoyable. Our heartfelt gratitude extends to the esteemed invited speakers, dedicated presenters, co-authors, engaged workshop attendees, invaluable resource personnel, and the diligent organisers and the participants of the mini-hackathon.

We extend heartfelt thanks to our sponsors, particularly our strategic partner – Octave: Data and Advanced Analytics Division of the John Keells Group, our Silver partner- Creative Software, Bronze partner Altria Consulting (Pvt) Ltd and Rootcode Labs, the workshop partner, for their generous financial support.

Organizing a conference of this magnitude is a collaborative effort. We express our gratitude to the Vice-Chancellor of the University of Colombo, Senior Professor (Chair) H.D. Karunaratne, for his wisdom and guidance. Special thanks also to the Dean of the Faculty of Science, Senior Professor (Chair of Physics) Upul Sonnadara, for continuous and invaluable support. We would like to acknowledge the Board of Management of the Center for Data Science for their encouragement and blessings. Lastly, a sincere thank you to the Head of the Department of Statistics, Dr. Rushan Abeygunawardhana, and all the dedicated staff members of the Department of Statistics for their untiring efforts to ensure the success of this conference.

Dr. J. H. D. S. P. Tissera	Dr. S. D. Viswakula
Co-chair/ ICDS 2023	Co-chair/ ICDS 2023
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 2023) organised by the Centre for Data Sciences and the Department of Statistics of the University of Colombo.

At a time where AI is transforming the world, I am delighted that the Center for Data Science has chosen the theme ‘’Data Science in the Age of Artificial Intelligence‘’ for the ICDS2023, where eminent speakers deliver talks covering multiple topics. Pre-conference workshops and a mini hackathon for students are also a few highlights of the event.

I congratulate the Director of the Center and the Head and all members of staff of the Department of Statistics for continuously contributing towards the development of the field of Data Science, in the Faculty of Science, in all possible ways. In particular, the hard work of academics, which has gone into establishing the Center and to develop both infrastructure and academic programs for undergraduates in relation to his field, is duly commendable.

While extending my warmest wishes to all the speakers, organisers and the participants of this conference, I hope this conference will produce stronger academic and industry networks that would provide better and bigger opportunities for both undergraduates and the postgraduates of the Faculty of Science to pursue a career in the field of Data Science.

Senior Professor H D Karunaratne
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 2nd International Conference on Data Science 2023, '*Data Science in the Age of Artificial Intelligence*'. This conference was first conceptualized and organised 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 at the University of Colombo was a very special occasion for all of us since we celebrated 100 years of excellence in science education that year. This year's conference, ICDS 2023, which is timely due to the rapid expansion of AI, offers an excellent platform for the exchange of scientific and technical knowledge as well as information related to the emerging field of Data Science across many application areas. With conference tracks ranging from machine learning to responsible AI and with experts coming together from industry and academia, we expect knowledge sharing to be at a high level at this conference.

The ICDS 2023 conference includes an inaugural session with keynote addresses and technical sessions organised in 8 tracks to present research papers and pre-conference workshops focused on AI and machine 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 organising committee for making 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 2023 (ICDS 2023) is one of the main events jointly organised by the Center for Data Science and the Department of Statistics of the University of Colombo, biannually. This year, the conference is organised for the second time with the timely theme of “Data Science in the Age of Artificial Intelligence”. This conference provides the opportunity for data scientists to present their novel concepts and applications of data science techniques to solve real world problems. This conference will help to enhance the collaboration between academia and industry. ICDS 2023 includes keynote speeches, several contributed sessions, two workshops and a mini hackathon. I trust that the ICDS 2023 will help to share new developments and enhance the knowledge in the field of data science.

I take this opportunity to extend my heartfelt gratitude to the multitude of individuals whose unwavering dedication has been instrumental in making this year's conference a reality. I extend my sincere thanks to the conference delegates, secretaries, track chairs, track coordinators, session chairs, the panel of reviewers, editors, members of the organising committee, and plenary speakers for their invaluable contributions. I am also grateful for the support provided by the Vice-Chancellor of the University of Colombo, Dean of the Faculty of the Science and my colleagues at the Department of Statistics, University of Colombo, whose contributions have been vital to the success of this conference.

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

Keynote Speaker

Professor Saman Halgamuge, Fellow of IEEE, IET and AAIA, received the B.Sc. Engineering degree in Electronics and Telecommunication from the University of Moratuwa, Sri Lanka, and the Dipl.-Ing and Ph.D. degrees in data engineering from the Technical University of Darmstadt, Germany. He is currently a Professor of the Department of Mechanical Engineering of the School of Electrical Mechanical and Infrastructure Engineering, The University of Melbourne (UoM). He served as a Distinguished Visiting Professor of Universities in Singapore, Malaysia, China, Sri Lanka and Indonesia and also held Prof V. K. Samaranayake endowed visiting professorship at UCSC, University of Colombo. He is listed as a top 2% most cited researcher for AI and Image Processing in the Stanford database. He was a distinguished Lecturer of IEEE Computational Intelligence Society (2018-21). He supervised 50 PhD students and 16 postdocs in Australia to completion. His research is funded by Australian Research Council, National Health and Medical Research Council, US DoD Biomedical Research program and International industry. His previous leadership roles include Head, School of Engineering at ANU and Associate Dean of the Engineering and IT school of UoM.

Keynote Speaker

David Hunter is a professor of statistics at Penn State University and has directed Penn State's newly-formed AI Hub since March 2023. He served as head of the Department of Statistics from 2012 to 2018 and has been heavily involved in developing Penn State's data science programs. He is a fellow of both the American Statistical Association and the Institute for Mathematical Statistics and is best known for work in three areas: Statistical algorithms, most notably the class known as MM algorithms; statistical modeling of networks; and mixture models, particularly finite mixtures in which the component distributions are not parametrically specified.

Keynote Speaker

Dr Rajitha Navarathna, a former Imagineer at Walt Disney Imagineering within The Walt Disney Company in the USA, currently serves as the Principal Data Scientist at OCTAVE, the John Keells Group Centre of Excellence for Big Data analytics. Prior to this role, he led AI teams at 99x and was the Lead Data Scientist at MIT.

From 2012 to 2019, Rajitha’s impactful career extended as an Imagineer at The Walt Disney Imagineering. Within the Research and Development team, he drove innovation through research publications and patented breakthroughs. In 2015/2016, his research contributed to the creation of a comprehensive data science lab at ABC media networks, focusing on understanding audience behavior. Rajitha’s research interests encompass computer vision, applied machine learning, facial tracking, facial expression recognition, and modeling machine vision systems capable of interpreting the world through extensive vision data. He earned a BSc Engineering degree with First Class Honors from the University of Peradeniya in Sri Lanka and completed his PhD in computer vision at the Queensland University of Technology in Australia in 2013.

Rajitha’s global impact spans collaborations with organizations like CSIRO, QUT, AutoCRC in Australia, Walt Disney Imagineering in Los Angeles and Pittsburgh, ABC media networks in Seattle, USA and USA, Disney Research Zurich in Switzerland. He has published around 30 academic papers and holds several US patents.

Beyond his industrial accomplishments, Rajitha is deeply passionate about mentoring research students. He has dedicated his time to guiding research students from various universities in Sri Lanka, as well as masters and PhD students from the USA, Australia, Canada, and Switzerland.

Keynote Speaker

Devini Senaratna is a Data Scientist with over a decade of experience with industry giants, Meta (formerly Facebook) and Booking.com, as well as successful stints in two machine learning startups. She has held several senior positions, managing as well as individually contributing and has worked across three continents: North America, Asia and Europe. Her expertise spans the areas of machine learning, causal inference, experimentation and natural language understanding. Devini holds a BSc in Statistics from the University of Colombo, Sri Lanka and a MSc in Statistics from Stanford University, USA. She has earned multiple awards and accolades including the Gold Medal for the Most Outstanding Science Student of 2010, University of Colombo and the Exceptional Global Talent Endorsement by TechNation, UK in 2023.

International Conference in Data Science 2023 (ICDS 2023)

19th - 20th December 2023, University of Colombo, Sri Lanka

Conference Committees

Advisory Committee

- **Senior Professor H.D. Karunaratne,**

  Vice-Chancellor, University of Colombo, Sri Lanka
- **Senior Professor D.U.J. Sonnadara,**

  Dean, Faculty of Science, University of Colombo, Sri Lanka
- **Dr. R.A.B. Abeygunawardana,**

  Head, Department of Statistics, University of Colombo, Sri Lanka
- **Dr. A.R. Weerasinghe,**

  Senior Lecturer, University of Colombo - School of Computing, University of Colombo, Sri Lanka

Conference Co-Chairs

Dr. J. H. D. S. P. Tissera

Dr. S. D. Viswakula

Conference Secretary

Dr. G.A.C.N. Priyadarshani

Review Committee

Dr. G. P. Lakraj (**Chair**)

Dr. J. H. D. S. P. Tissera

Dr. S. D. Viswakula

Dr. G.A.C.N. Priyadarshani

Workshop Committee

Dr. R. V. Jayatillake **(Chair)**

Dr. I. T. Jayamanne

Mr. O.N.S. Senaweera

Academic Support Staff – Mr. R.S.A.U. Dharmarathna, Mr. D.S. Ruwankumara, Ms. S.P.P.M. Sudasinghe

Programme Organising Committee

Mr. E. R. A. D. Bandara **(Chair)**

Dr. J. H. D. S. P. Tissera

Dr. A. A. Sunethra

Dr. K. A. D. Deshani

Dr. G. H. S. Karunarathna

Academic Support Staff – Ms. K.U.S. Kumarathunga, Mr. D.B.S. Suwaris, Ms. H.M.S. Yasara, Ms. S.P.P.M. Sudasinghe, Mr. R.S.A.U. Dharmarathna, Ms. K.G.N. Senanayaka, Ms. W.M.D. Nawanjana, Ms. Y. Dilaxiha, Ms. D.H.D.N. Hettige, Ms. M.D.T.U. Ranasinghe, Ms. M. K. D. Tharushika

Mini Hackathon Organising Committee

Dr. S. D. Viswakula **(Chair)**

Dr. J. H. D. S. P. Tissera

Dr. G. P. Lakraj

Mr. O. N. S. Senaweera

Dr. K. A. D. Deshani

Dr. A. A. Sunethra

Student Organisers – Ms. Thiruni Withana, Mr. Yasas Jayaweera, Mr. Susara Ouchithya, Ms. Sanuji Devasurendra, Mr. Anjana Jayasinghe, Mr. Sahan Madusanka, Ms. Kaushalya Atthanayake, Ms. Sehara Sooriyarachchi, Mr. Disura Chandrasekara, Mr. Buddhima Senaratne, Ms. Pramudi Rajamanthri, Ms. Neelya

Jayasundara, Ms. Lasani Balasuriya, Mr. Kavishka Palihena, Mr. Ravindu Nishal, Mr. Ruwinda Rowel, Ms. Darshika Wijesena, Ms. Darshi Yashodha, Ms. Hashini Nimesha, Ms. Kaveesha Vidushinie, Mr. Nayana Chathuranga, Mr. Olindu Eranja, Ms. Senuri Perera, Ms. Vivada Lokugamage, Ms. Yashika Asiriwardhana, Ms. Malsha Kavindi, Ms. Nipuni Sandunika, Ms. Ranushi Nimthara, Ms. Ravindi Sandurashmi, Mr. Bhashitha Wijesinghe, Ms. Tashini Ramindi

Administrative and Support Staff

Mr. N.D Suduwella

Mr. H. K. T. Nanayakkara

Ms. R.M.N.E.K. Rathnayake

Ms. R. A. K. Kithmini

Mr. W. D. M. C. Withanage

Panel of Reviewers

All abstracts included in the Proceedings of the International Conference in Data Science 2023 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. R. A. B. Abeygunawardana**

  Department of Statistics, University of Colombo
- **Mr. Dinesh Asanka**

  Department of Industrial Management, University of Kelaniya
- **Mr. E. R. A. D. Bandara**

  Department of Statistics, University of Colombo
- **Prof. Vasana Chandrasekara**

  Department of Statistics and Computer Science, University of Kelaniya
- **Dr. Dilanka Shenal Dedduwakumara**

  School of Mathematical Sciences, University of Adelaide, Australia
- **Dr. Mahasen Dehideniya**

  Department of Statistics and Computer Science, University of Peradeniya
- **Dr. K. A. D. Deshani**

  Department of Statistics, University of Colombo
- **Dr. Kushani P. De Silva**

  Department of Mathematics, University of Colombo
- **Dr. Anurika P. De Silva**

  Melbourne School of Population and Global Health, University of Melbourne, Australia
- **Dr. H. A. S. G. Dharmarathne**

  Department of Statistics, University of Colombo
- **Dr. Pansujee Dissanayake**

  Department of Mathematics and Statistics, University of Kelaniya
- **Dr. Neluka Devpura**

  Department of Statistics, University of Sri Jayewardenepura

- **Dr. Niroshinie Fernando**

  School of Info Technology, Deakin University, Australia
- **Dr. G. P. Lakraj**

  Department of Statistics, University of Colombo
- **Dr. Gayan Hettiarachchi**

  Department of Physics, Osaka University, Japan
- **Dr. Budditha Hettige**

  Department of Computer Engineering, General Sir John Kotelawala Defense University
- **Dr. Anuradha Hewaarachchi**

  Department of Statistics and Computer Science, University of Kelaniya
- **Dr. Isuru Udayangani Hewapathirana**

  Software Engineering Teaching Unit, University of Kelaniya
- **Dr. I. T. Jayamanne**

  Department of Statistics, University of Colombo
- **Dr. R. V. Jayatillake**

  Department of Statistics, University of Colombo
- **Dr. J.A. Jeewanie**

  Department of Computer Science, University of Ruhuna
- **Dr. Pradeep Kalansooriya**

  Department of Computer Science, General Sir John Kotelawala Defense University
- **Dr. Prasanna Karunanayaka**

  Department of Radiology, Department of Public Health Sciences, Department of Neural and Behavioral Sciences, Penn State Neuroscience Institute, Penn State University, USA
- **Dr. G. H. S. Karunarathna**

  Department of Statistics, University of Colombo
- **Dr. C. I. Keppitiyagama**

  University of Colombo School of Computing

- **Mr. Yasiru R. Kirindearachchi**

  School of Computing, Engineering and Mathematics, Western Sydney University, Australia
- **Dr. M. W. P. Maduranga**

  Department of Computer Engineering, General Sir John Kotelawala Defense University
- **Dr. C. H. Magalla**

  Department of Statistics, University of Colombo
- **Dr. Thilini V. Mahanama**

  Department of Industrial Management, University of Kelaniya
- **Dr. G. A. C. N. Priyadarshani**

  Department of Statistics, University of Colombo
- **Dr. Rasika Rajapaksha**

  Faculty of Computing and Technology, University of Kelaniya
- **Mr. O. N. S. Senaweera**

  Department of Statistics, University of Colombo
- **Dr. Edirisuriya M. D. Siriwardane**

  Department of Physics, University of Colombo
- **Dr. A. A. Sunethra**

  Department of Statistics, University of Colombo
- **Dr. Priyanga Dilini Talagala**

  Department of Computational Mathematics, University of Moratuwa
- **Dr. Thiyanga S. Talagala**

  Department of Statistics, University of Sri Jayewardenepura
- **Dr. J.H.D.S.P. Tissera**

  Department of Statistics, University of Colombo
- **Dr. Hakim Usoof**

  Department of Statistics and Computer Science, University of Peradeniya
- **Dr. S. D. Viswakula**

  Department of Statistics, University of Colombo

- **Dr. Dilani Wickramaarachchi**

  Department of Industrial Management, University of Kelaniya
- **Ms. D. S. Wickramarachchi**

  Department of Statistics, University of Colombo
- **Dr. Lakmini K. N. Wijesekara**

  School of Computing, Engineering and Mathematics, Western Sydney University, Australia
- **Dr. Rushani Wijesuriya**

  Murdoch Children's Research Institute, Melbourne, Australia

Table of Contents

Title and Author/s	Page No:
Machine Learning for Longitudinal Studies in Medical Research **(Keynote speech)** **Prof. Saman Halgamuge**	1
In Defense of Interpretable Models **(Keynote speech)** **Prof. David Hunter**	2
Advancing Data Science Proficiency: Next-Gen Data Science Proficiency and Decision-Making Excellence in the Digital Age **(Keynote speech)** **Dr. Rajitha Navarathna**	3
Bridging Worlds: Navigating the Interplay of Causal Inference and Generative AI **(Keynote speech)** **Ms. Devini Senaratna**	4
Technical Sessions	
Comparing CT Scan Images of Lungs **Dhammika Amaratunga**	6
A Comparative Study of Academic Performance between Undergraduate Athletes and Non-Athletes of the Faculty of Science, University of Colombo using Machine Learning Techniques **E A I P De Silva, G H S Karunarathna**	7
Deep Learning Model for Lip Synchronization in Sinhala Language **D M P Disanayaka, A R Weerasinghe**	8
DR-Vision: A Vision Transformer Model for Diabetic Retinopathy Classification using Fundus Images **P Gajithira, M G A Saumyamala, K A D Deshani**	9
Ensemble-Based Classification of COVID-19 Radiographic Images using Spatial, Textural, and Intensity-Based Features **Susmita Ghosh, Abhiroop Chatterjee**	10

Title and Author/s	**Page No:**
IoT Based Smart Paddy Field Monitoring & Water Management System **H H D Malithi Pramoda Halpandeniya, Samadhi Rathnayake, Amali Gunasinghe**	11
Developing a Forecasting Model for the Exports of Ceylon Tea **B L I Kalpani, C H Magalla**	12
Development of a Forecasting Model to Determine the Interest Rate in Sri Lanka – An Approach Based on the 91-Day Treasury Bill Rate **T J Kotelawala, S D Viswakula**	13
Analysing The Nuances of Supermarket Customer Behaviour for Enhanced Retail Marketing Strategies: A Case Study on Customer Purchases **T S M A Muthalib, K A D Deshani, O N S Senaweera**	14
A Comparative Study of Deep Learning Approaches with Vision Transformers for Multi-Class Dermoscopic Image Classification W M D Nawanjana, G P Lakraj	15
Application of Feature Selection and Machine Learning for Non-invasive Breast Cancer Prediction Using Circulating MicroRNAs **G P Paranawithana**	16
An Evolutionary Keystroke Dynamics-Based Stress Detection System through an Incremental Learning Based Approach for IT Professionals **M. S. D Perera, S. V. Bartholomeusz, H. M. Samadhi Chathuranga Rathnayake, Devanshi Ganegoda**	17
A Deep Learning Based Approach for Car Insurance Claim Management **B K A S Rodrigo, C H Magalla, Prasad Wimalaratne**	18
Paddy Crop Quality Monitoring and Mapping Application using Object Detection Techniques **W D Nilakshi Sandeepanie, Samadhi Rathnayake, Amali Gunasinghe**	19

Title and Author/s	**Page No:**
A Machine Learning Approach for Predicting the Depression Status Among Undergraduates **D S Sonnadara, S D Viswakula, M D T Attygalle**	20
Predicting the Outcome of T20 International Matches Based on Historical Statistics and Crowd Opinions **S P P M Sudasinghe, S D Viswakula, G P Lakraj**	21
The Study on Malaria Cases Diagnosed in Sri Lanka after Returning from Malaria-Endemic Countries **B D S Suwaris, J H D S P Tissera, Pubudu Chulasiri**	22
Convolutional Neural Network and Vision Transformer Models with Dynamic Attention for Emotion Recognition In Facial Dynamics-Based Stress Detection **Janudi Ranasinghe, Amantha Jayathilake, Samadhi Rathnayake, Devanshi Ganegoda**	23
Comparative Analysis of Machine Learning Algorithms for Breast Cancer Tumor Grade Prediction **M D T U Ranasinghe, J H D S P Tissera**	24
Suspicious Activity Likelihood Detection, Indication & Tracking, and Alert System to Prevent Crimes **E Sankeetha, T S W Walagedara , A J O Geenath , G R S D Gunasinghe, D S Alwis**	25
Identifying the Frequently Occurring Aquatic Insect Assemblages Using Association Rule Mining Techniques **K N W Tennakoon, M D T Attygalle, P K T N S Pallewatta, M S Kanakarathna**	26
A Study on Mental Disorders Among Psychiatric Patients in Kalutara District: A Hospital-Based Study **R A C Tharundi, A A Sunethra**	27

Title and Author/s	**Page No:**
Using Causal Inference in Portfolio Optimization: A Case Study **S I Ubayawickrema, K M M H Siriwardana, S D Viswakula**	28
Feature Engineering and Music Genre Classification: A Showdown between XGBoost and Neural Networks **W M H G T C K Weerakoon, M S H Peiris, P B S N Ariyathilake**	29
Skin cancer image classification using Vision Transformer network **W M D D B Wijesundara, C T Wannige, M K S Madushika, P L A N Liyanage**	30

Machine Learning for Longitudinal Studies in Medical Research

Prof. Saman Halgamuge

Professor of the Department of Mechanical Engineering, School of Electrical Mechanical and Infrastructure Engineering, University of Melbourne

This talk addresses key deficiencies of Machine Learning stopping them from using for continuous learning as necessary in longitudinal experimental studies or in a pandemic when important data arrives over a long period of time, and we need to be able to make decisions based on data at hand. New methods that help continuous learning are briefly introduced and their applications in medical research are discussed. Unlike a static machine learning model that uses all the data at the end of an experiment, these methods generate a progression trajectory thus far at each sampling timepoint of the experiment. Results on simulated data, for which the true progression trajectories are known, verified the ability to capture and visualize the trajectories accurately and relative to each other. I will refer to several on-going projects in my lab funded by Australian and US grants.

In Defense of Interpretable Models

Prof. David Hunter
Professor of Statistics, Penn State University

The most complicated regression models used for mimicking human decision-making these days are said to have more than one trillion parameters. While the predictive capabilities of such models are impressive and helpful for certain tasks, they are ill-suited for certain tasks for multiple reasons. This talk describes a true story of a simplistic regression model, now several decades old, whose interpretability was the hallmark of its usefulness. The United States Supreme Court recently decided a case on the use of race in university admissions. This case has a legal precedent, decided in 2003 in the case of Gratz v. Bollinger, which hinged on interpreting the coefficients in a logistic regression model. This talk briefly discusses the history of the legal case, arguing that while this model is ultra-simplistic by modern standards, it enabled a debate which would not have been possible using a more complicated predictive approach. As data scientists, we do not always benefit society most by striving for the best possible predictive performance.

Advancing Data Science Proficiency: Next-Gen Data Science Proficiency and Decision-Making Excellence in the Digital Age

Dr. Rajitha Navarathna
Principal Data Scientist, OCTAVE, John Keells Group

As data science continues to evolve, discovering the next-level skills vital for the future becomes increasingly important. Data science plays a central role in making informed decisions across various business sectors, demanding a diverse set of abilities to navigate the complexities of the digital age. This includes mastering advanced experimentation and statistical modeling, domain-specific expertise, and incorporating ethical considerations.

In the latter part of the talk, we narrow our focus to the critical aspect of decision-making in data science, highlighting innovative strategies tailored to making intelligent decisions to address real-world analytical challenges. By leveraging statistical modeling and optimization techniques, we can delve into effective approaches that bridge the gap between theory and application, ensuring relevance and success in solving complex problems in the rapidly changing realm of data science. It's all about finding smart ways to make decisions that actually work in the real world.

Bridging Worlds: Navigating the Interplay of Causal Inference and Generative AI

Ms. Devini Senaratna

Senior Data Scientist, Meta (Formerly Facebook), London

The application of experimental and observational causal inference has garnered increased academic and commercial interest. Concurrently, the field of generative artificial intelligence (Gen AI) has experienced exponential growth following the releases from OpenAI. While the primary objective of causal inference is to evaluate cause-and-effect relationships, generative AI is designed to "generate" new data that emulates learned distributions. Consequently, causal inference plays a pivotal role in planning and reflection, whereas generative AI excels in more creative tasks such as Q&A or image generation.

This discourse discusses the ongoing topic surrounding the relationship between generative AI and causal inference, exploring two perspectives. Firstly, it delves into how Gen AI can enhance the field of causal inference by addressing challenges such as generating counterfactuals, mitigating data scarcity issues and enhancing model interpretability. The second component delves into the ethical considerations associated with Gen AI, including the risk of scientific fraud within academia and industry, as well as the nuanced definition of intellectual property.

ICDS2023 - Technical Sessions

Comparing CT Scan Images of Lungs

Dhammika Amaratunga*

Independent consultant, Colombo, Sri Lanka

*damaratung@yahoo.com

Lung diseases such as emphysema are associated with significant structural abnormalities which limit airflow into and out of the lungs, greatly impeding proper functioning of the lungs. During the development of an experimental procedure that would be capable of characterizing structural abnormalities with high sensitivity through the use of high-resolution imaging, CT scan images of lungs were taken on 18 mice. Of the 18, 6 were healthy, 6 had moderate emphysema, and 6 had severe emphysema. The voxel intensity data of the images generated by these scans were analyzed by comparing their distributions. This was accomplished by calculating pairwise dissimilarities between distributions and applying an F test using these dissimilarities. In addition, a trend test was applied on a one-dimensional representation of the dissimilarities to study the effect of increasing disease severity. Other supplementary analyses were also done. All analyses indicated a statistically significant shift towards lower intensities in the diseased mice. In addition, the use of different dissimilarities for this type of data was explored via a simulation. The study and the simulation showed the strong effectiveness of this approach for the study of lung disease.

Keywords: *Voxel intensities, dissimilarities, permutation test*

A Comparative Study of Academic Performance between Undergraduate Athletes and Non-Athletes of the Faculty of Science, University of Colombo using Machine Learning Techniques

E A I P De Silva[1]*, G H S Karunarathna[1]

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

*edirisingheip@gmail.com

The involvement of sports activities is a possible factor associated with students' academic performance. This aspect is primarily associated with university undergraduates who wish to participate in athletics but are unsure how it might be associated with their academic performance. Therefore, the objectives of the study were to identify factors associated with the academic performance of the students, and factors that lead university students to engage in sports. This study was conducted through a survey with a sample of 271 students out of a target population of 2225 using a stratified random sampling technique, based on undergraduates' academic year. The explanatory analysis revealed a substantial association in the GPA of level IV undergraduate athletes. Moreover, the number of sports engaged in, exhibited a significant association with an undergraduate's GPA, and the amount of time spent on both academic and athletic pursuits showed a notable association. In this study, 7 machine learning techniques, namely, logistic regression, KNN, linear discriminant analysis, random forest regression, support vector regression, gradient boosting, and decision trees were used to find out the best fitted model through performance indices. Random forest classification emerged as the most effective model in determining the factors associated with undergraduates' choice to participate in sports, with 76% test accuracy. Notably, the student's academic year was identified as a key determinant for engaging in sports activities at the university from the random forest classification. Furthermore, through a comparative analysis of undergraduates' perspectives, it was shown that sports were typically seen as having a positive association with academic pursuits and improving overall academic performance. Conversely, non-athletes expressed concerns about negative implications, such as time constraints on academic achievements.

Keywords: *Comparative study, Athletes, Academic performance, Machine learning*

Deep Learning Model for Lip Synchronization in Sinhala Language

D M P Disanayaka[1]*, A R Weerasinghe[2]

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

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

*pulindisanayaka@gmail.com

In the captivating world of animation, where characters spring to life and narratives unfold through the art of visual storytelling, achieving a seamless marriage between auditory and visual elements becomes a paramount pursuit. One of the key processes in this procedure is lip synchronization, which is an intricate task done by experienced artists in the animation industry with hours of work to bring the best combination of auditory and visual elements to the audience. While this technique has been extensively explored for widely spoken languages like English, Hindi, etc., its application for languages with narrower representation like Sinhala is very low. Creating a system that can manipulate this intricate task is the main objective of this research. While previous researchers have tried rule-based algorithms to match "phonemes" which are fundamental sounds formed in any language to "visemes" a visual representation of lip movement, in this research the authors tried to match the phonemes to a face mesh with 468 points which gives a greater degree of flexibility when it comes to generating the final facial animation rather than using static predefined "visemes". The selected deep-learning approach was used to predict the 3D positioning of 468 facial landmarks. After initial training and testing using different architectures and activation layers, the most promising results were identified when Long Short-Term Memory layers were incorporated into the model, proving particularly effective in capturing the coarticulation phenomenon observed in speech. The final model achieved a validation Root Mean Square Error of 0.1109. The model's effectiveness hinges on the accuracy of the output it produces, making subjective assessment a crucial factor in gauging its performance. The model achieved an average rating of 68.50% and 70.60% in a survey done for short and long speeches respectively.

Keywords: *Natural Language Processing, Deep Learning, Lip Synchronization, Linguistics*

DR-Vision: A Vision Transformer Model for Diabetic Retinopathy Classification using Fundus Images

P Gajithira[1]*, M G A Saumyamala[1], K A D Deshani[1]

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

*gajipuvi@gmail.com

Diabetic Retinopathy (DR) is a critical retinal complication associated with prolonged diabetes, often leading to vision impairment and blindness. Its progression is categorized into five stages, ranging from No DR (Stage 0) to Proliferative DR (Stage 4). Early detection plays a pivotal role in effective treatment and preventing disease progression. However, conventional screening methods are hindered by low patient turnout, delayed diagnosis, and high costs. To address these challenges, the adoption of automated detection systems, predominantly reliant on Convolutional Neural Networks (CNNs), has emerged. However, CNN architectures struggle to capture the global context of images. This research introduces an approach utilizing the Vision Transformer (ViT) model, which employs a multi-head self-attention mechanism to capture global contextual information for DR classification and visualizing activation regions for decision-making. The study systematically evaluates ViT models for DR classification tasks, comparing Transfer Learning (TL) with training from scratch, assessing model robustness using an external dataset, and providing insights into model predictions from a medical expert's perspective. Utilizing the EYEPACs dataset known for its richness and variation, various TL methods, including Feature Extraction (FE) and FE with fine-tuning (FE + FT), are applied to ViT models such as ViT B/16 (ViT Base with 16 layers), ViT B/32 (ViT Base with 32 layers), and ViT R26+S32 (Hybrid model of ResNet with 26 layers and ViT Small with 32 layers). The ViT B/16 FE+FT model demonstrates outstanding performance, achieving an accuracy of 96.27%, specificity of 99.06%, sensitivity of 99.51%, and a quadratic kappa score of 0.958, surpassing the ResNet50 CNN architecture. Named DR Vision, this model exhibits superior stage-wise performance, albeit with some limitations in identifying stages 3 (Severe Non-Proliferative DR) and 4 (Proliferative DR). The model's generalizability assessed using the IDRiD dataset, emphasizes the necessity for improvement. The visualization of explainable regions within DR Vision, aligning reasonably well with a medical expert's observations, substantiates the model's potential in DR classification and transparency in the model’s decision-making process leading to potential clinical utility.

Keywords: *Vision Transformer, Diabetic Retinopathy, Attention models, Explainable AI*

Ensemble-Based Classification of COVID-19 Radiographic Images using Spatial, Textural, and Intensity-Based Features

Susmita Ghosh[1], Abhiroop Chatterjee[1*]

[1]Jadavpur University, Kolkata, India

*abhiroopchat1998@gmail.com

The COVID-19 pandemic has underscored the critical need for precise diagnostic tools. Radiographic imaging, particularly chest X-rays and CT scans, has emerged as a vital modality for detecting COVID-19-related lung abnormalities. Several machine learning techniques exist to diagnose Covid-19 from such imaging modalities. In this article, an innovative approach is introduced using multimodal features extracted from CT scan images and ensemble learning to enhance the COVID-19 diagnosis. The dataset used here encompasses a wide range of COVID-19-positive cases and non-COVID-19 cases. To improve image quality and information extraction, preprocessing techniques like resizing, Gaussian blur, and dilation are employed. These steps enhance the ability to capture informative image features. Intensity-based, spatial, and textural features have been exploited for enhanced classification. Intensity-based features capture statistical properties of pixel intensities, while spatial features quantify geometric characteristics of image regions, and texture features leverage the Gray-Level Co-occurrence Matrix (GLCM) to encompass essential texture attributes of the images. The classification strategy used in this paper revolves around ensemble learning, utilizing three diverse base classifiers: Random Forest, Support Vector Machine, and Light Gradient Boosting Machine, each with unique strengths in image classification. These base classifiers are trained on standardized feature vectors from the training dataset, and their predictions are aggregated through averaging the decisions. This ensemble approach yields a robust model for COVID-19 image classification, achieving an overall test accuracy of 92.42%. The performance of the ensemble model is assessed using a dedicated test dataset, demonstrating its superiority (in terms of accuracy) over other methods that utilize hand crafted as well as deep features. Moreover, the potential clinical implications of our research and outline avenues for future enhancements are also discussed.

Keywords: *Ensemble Learning, Gray-Level Co-occurrence Matrix, Random Forest Classifier, Support Vector Classifier (SVC), Light GBM Classifier.*

IoT Based Smart Paddy Field Monitoring & Water Management System

H H D Malithi Pramoda Halpandeniya[1]*, Samadhi Rathnayake[1], Amali Gunasinghe[1]

[1]Department of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

*malithihalpandeniya@gmail.com

In the context of advancing agricultural practices, the imperative for sustainable and efficient irrigation solutions has grown significantly. This research addresses this demand by introducing a tailored smart irrigation system designed specifically for paddy fields. The primary focus of the study is on managing optimal water levels critical for successful paddy cultivation—a challenge traditionally complicated by the dynamic nature of soil moisture. The research involves the implementation of a large-scale smart irrigation system in an industrial paddy cultivation area. An Alternate Wetting and Drying (AWD) system is employed for water supply. In contrast to conventional methods where water is screened by a dam and distributed based on general requirements, our system utilizes real-time soil moisture readings from designated lanes, providing a more accurate assessment of the actual water needs. A custom device, equipped with sensors measuring environmental parameters such as temperature, humidity, rainfall, and soil moisture, is strategically deployed across the field. These devices form an interconnected network, collecting data and generating average values for predicting required water supply. The prediction model employs the ARIMA (Autoregressive Integrated Moving Average) methodology, enhancing the precision of water supply forecasts. To facilitate user access to data, a web application is developed. Farmers can input the device ID corresponding to their location, obtaining comprehensive information about environmental conditions and the recommended water supply. This integrated approach leverages IoT technology to enhance the efficiency, accuracy, and sustainability of paddy field irrigation, contributing to the advancement of precision agriculture.

Keywords: *Internet of Things, climate measurements, water management, AWD method, sensors, rice cultivation*

Developing a Forecasting Model for the Exports of Ceylon Tea

B L I Kalpani[1]*, C H Magalla[1]

[1]Department of Statistics, University of Colombo, Colombo 03, Sri Lanka

*isharabaddeliyanage@gmail.com

The tea industry in Sri Lanka plays an essential role in terms of its contribution to national output. Since the global market demand for Ceylon tea changes over time and becomes competitive, forecasting for tea export could have a critical impact on the future Sri Lankan economy. This study aims at developing a model to forecast the monthly Ceylon tea export volume considering all types (Bulk Tea, Packeted Tea, Tea Bags, Other) of tea exported. Monthly time series data for production, Free-On-Board price, Colombo auction price, quantity sold at Colombo auction, import volume, GDP, inflation rate, exchange rate, average weighted prime lending rate, rainfall, average temperature, day-time relative humidity and night-time relative humidity for a period of 21 years (2001-2021) were used in the analysis. These historical data were used to identify the components using ACF (autocorrelation function) plots, PACF (partial autocorrelation function) plots and CCF (cross- correlation function) plots and the necessary variables were adjusted using stationarity tests in order to forecast export volume of Ceylon tea. Out of a variety of traditional statistical and machine learning forecasting techniques, Feed Forward Neural Network architecture with one hidden layer model under Artificial Neural Networks, along with an appropriate set of 89 features using XGBoost feature selection, was identified as the optimal model to forecast the monthly Ceylon tea export volume with RMSE and MAE values of 0.1045 and 0.0788, respectively. The study recommends that, machine learning algorithms tend to be more accurate when forecasting the monthly export volume of Ceylon tea compared to traditional models. Based on the results, Free-On-Board Price, Colombo auction price and GDP were the only non-significant features in forecasting monthly Ceylon tea exports.

Keywords: *Forecasting, Machine learning, Tea export, Sri Lanka*

Development of a Forecasting Model to Determine the Interest Rate in Sri Lanka – An Approach Based on the 91-Day Treasury Bill Rate

T J Kotelawala[1]*, S D Viswakula[2]

[1] Informatics Institute of Technology (IIT), Colombo, Sri Lanka

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

*theodore.kotelawala@gmail.com

The increased level of economic activity, together with the increased number of economic agents have shown that accurately forecasting interest rates, especially in a volatile economic environment, is essential. A country needs to identify the specific determinants of their interest rates and hence, arrive at a forecasting model that can accurately determine the interest rate. Considering the change in the exchange rate regime in 2001 (and the significant relationship between exchange rates and interest rates as specified in certain studies) and the higher level of transactions with the global economy after the end of the civil war in 2009, a suitable model must be identified to forecast the interest rate in Sri Lanka. In the Sri Lankan context, a few studies have been conducted to explore the relationship between certain economic variables and interest rates. However, these studies have not proceeded to forecast interest rates based on the identified determinants. Only a single study was identified which had attempted to build a forecasting model to determine the interest rate, using the Sri Lanka Interbank Offered Rate (SLIBOR) as the proxy variable. However, with the discontinuation of "the compilation and publication of SLIBOR", the findings of this study were made invalid and hence a greater need exist interest rate forecasting based on a more relevant proxy. In this study, the 91-day Treasury Bill rate is used as a proxy for the interest rates owing to the high level of liquidity in both primary and secondary markets. The study follows an eclectic approach based on the sample period from 2000 to 2020 and data at monthly frequency is used. This data is used for the Long Short-Term Memory (LSTM) Network and Vector Autoregression with Exogenous Variables (VARX) models with 65% of the data as training data and 35% as testing data, to capture a higher period of data in testing model accuracy while providing a sufficient data range for the model to be trained on. It is identified that the LSTM model with a dropout value of 0.05, performs best based on the values of the calculated error metrics.

***Keywords**: Interest Rate, 91-day Treasury Bill Rate, Long Short-Term Memory (LSTM) Network, Vector Autoregression with Exogenous Variables (VARX)*

Analysing the Nuances of Supermarket Customer Behaviour for Enhanced Retail Marketing Strategies: A Case Study on Customer Purchases

T S M A Muthalib[1*], K A D Deshani[1], O N S Senaweera[1]
[1]Department of Statistics, University of Colombo, Sri Lanka
*thasneem.shehani@gmail.com

The customer is considered as the main stakeholder of any business, and meeting their needs is crucial for success. Currently supermarkets are gaining attention from customers due to the convenience of grocery shopping. However, it was found that the few available published work in this field have limited information regarding the techniques used and the methodology. This gives the prominence to perform a detailed market research, specifically in the Sri Lankan context. This study uses a multifaceted approach to explore supermarket customer behavior in Colombo Sri Lanka using two datasets, having product name, unit price, quantity and total amount as variables, with one containing 17,107 transaction details from 152 houses across 8 areas, and the other containing 22,623 with data of 5944 distinct items. Descriptive analysis revealed that supermarkets SM3 and SM2 (real names are labelled) are experiencing good sales, with vegetables and fruits on top, where rush days for grocery shopping are Saturdays or Thursdays and Fridays. Market Basket Analysis (MBA), through Apriori and FP-Tree algorithms which are association rule mining techniques, were used in this study to uncover relationships among objects. The study utilized K-means, Hierarchical clustering, and Latent Dirichlet allocation methods for consumer segmentation. The initial MBA uncovered rules for products, which can be used for future promotional discounts. Also 10 rules for product categories like beverages, biscuits, snacks, confectionaries, and personal care products were mined for planogram purposes. The method effectively enhanced customer engagement and provided insights into purchasing trends. Secondly, Hierarchical clustering average linkage method, with elbow plot and dendrograms were used to segment consumers into "Premium" and "Standard" based on their expenditure, revealing that Premium customers spend 3.5 times more than Standard customers. Consumers were characterized by their product choices using LDA method, based on best Silhouette score of 0.622, and were named as 'Balanced lifestyle seekers', 'Balanced diet consumers', 'Food explorers', 'Indulgence seekers' and 'Whole-food consumers'. Ultimately this study on the retail market aims to enhance customer satisfaction through customer behaviour dynamics.

***Key words**: Data mining, clustering, association rule mining, machine learning, market basket analysis*

A Comparative Study of Deep Learning Approaches with Vision Transformers for Multi-Class Dermoscopic Image Classification

W M D Nawanjana[1*], G P Lakraj[1]

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

*wmdnawanjana@gmail.com

Skin cancer is a serious global health concern, leading to the loss of many lives each year. Advanced melanoma, a severe form of skin cancer, often results in a life expectancy of less than five years. However, early detection significantly increases the chances of survival to over 95%. Due to time-consuming and error-prone clinical procedures, the identification and categorization of skin lesions are significantly impacted by the application of deep learning based techniques. Convolutional Neural Networks (CNNs) have been widely successful in image recognition tasks, but the recent introduction of Vision Transformers (ViTs) has brought new possibilities. This has sparked competition between CNNs and ViTs, pushing for advancements in computer vision strategies. The purpose of this study was to find a more accurate model for classifying multi-class dermoscopic skin lesion images. The research explored CNNs, ViTs, and their combined models with Transfer Learning. The study used the ISIC 2019 dataset from the International Skin Imaging Collaboration (ISIC). Different CNN models renowned for their proven performance in literature over the years, including ResNet50, Caption, MobileNetV3, EfficientNetV2, and ConvNeXt, were employed. Both feature extraction (FE) and feature extraction combined with fine-tuning techniques (FE+FT) were utilized in the study. The ViT models used in this study were ViT-B/16, ViT-B/32, ViT-L/16, and ViT-L/32. Among the ViT models, only feature extraction was employed due to resource limitations. The ensemble model training used the best fitted two models from CNN architectures, and one model from ViT architecture due to lack of computer resources. After evaluating both CNN FE and ViT FE models, it was determined that CNN models outperformed ViT models for the given assignment. Consequently, CNN FE + FT, were selected as the optimal choice and the best model was given by the ConvNeXt FE+FT model with an accuracy of 88.61%, precision of 89.0%, and specificity of 89%.

Keywords: *Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Transfer Learning, Fine-tuning, Image Classification*

Application of Feature Selection and Machine Learning for Non-invasive Breast Cancer Prediction Using Circulating MicroRNAs

G P Paranawithana[1]*

[1]Department of Mathematics, University of Ruhuna, Matara, Sri Lanka

*githmiparanawithana79@gmail.com

Breast cancer is a significant public health concern worldwide, including in Sri Lanka, where it is the most common cancer among women. Despite advances in medical technology, early detection remains a critical factor in improving patient outcomes. While mammogram screening has been widely utilised in breast cancer detection, its lower sensitivity to smaller tumours and radiation from mammography have been concerns for many years. Over the years, there has been significant interest in the development of non-invasive biomarkers for efficient diagnosis of breast cancer. Circulating microRNAs (miRNAs) are single-stranded, non-coding RNAs that regulate gene expression at the posttranscriptional level. miRNAs have demonstrated outstanding potential as a promising diagnostic tool for breast cancer due to their presence observed in breast tumours and accessibility through non-invasive techniques. However, the vast amount of miRNA data has posed the challenge of selecting the most informative features for making timely and accurate predictions. This study proposes an efficient machine learning framework combined with feature selection techniques for selecting the most significant and informative features for breast cancer prediction. This study evaluated the performance of several supervised machine learning methods, including Support Vector Machine, Decision Tree and Deep Neural Network using the GSE58606 dataset. The SMOTE technique was used to handle the class imbalance in the dataset. The deep learning classifier achieved an accuracy, F1-score, precision and recall of 99%, 98.4%, 98.5%, and 98.2%, respectively. Therefore, with further validation using data representing diverse populations, circulating microRNAs present a promising non-invasive diagnostic tool for early identification of breast cancer.

Keywords: *breast cancer, diagnosis, biomarkers, microRNAs, machine learning, feature selection*

An Evolutionary Keystroke Dynamics-Based Stress Detection System through an Incremental Learning Based Approach for IT Professionals

M. S. D Perera[1*], S. V. Bartholomeusz[1], H. M. Samadhi Chathuranga Rathnayake[1], Devanshi Ganegoda[1]

[1]Sri Lanka Institute of Information and Technology
Colombo, Sri Lanka
*it20020262@my.sliit.lk

This research paper presents a novel machine learning approach for real-time stress level detection through the analysis of individual keystroke dynamics. By capitalizing on the inherent uniqueness of typing patterns exhibited by each user, this methodology incorporates incremental learning to continually assimilate new user inputs, thereby enhancing the accuracy of the base model. A discreet Python programme seamlessly operates in the background, collecting keystroke dynamics without disrupting the user's experience. This unobtrusive data collection method distinguishes our work from prior studies that often relied on specialized keyboards, manufactured stressors, or physiological sensors. Central to our methodology is the hosting of the machine learning model on a Flask server, leveraging the versatility and practicality of web-based deployment. Powered by the Random Forest algorithm, our model showcases its efficiency in real-world applications, offering a continuous assessment of stress levels without intrusive measures. This research contributes a unique dimension to stress prediction, foregoing the need for external devices or artificial stress inductions. Moreover, it highlights the immense potential of machine learning and incremental learning-based paradigms in constructing adaptable, user-centric systems. As we look ahead, our future endeavours aim to fuse mobile phone touch keypress dynamics with keyboard data to create an aggregated predictive model, further enriching the comprehensiveness of stress assessment. In conclusion, this research underscores the transformative potential of technology in stress detection, advocating for unobtrusive yet robust methodologies. By blending seamlessly into users' interactions, our approach paves the way for a more holistic understanding of stress and opens avenues for its effective management in an increasingly technology-driven era.

Keywords: *keyboard, incremental learning, keystroke dynamics, random forest, stress detection, IT professionals*

A Deep Learning Based Approach for Car Insurance Claim Management

B K A S Rodrigo[1*], C H Magalla[1], Prasad Wimalaratne[2]

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

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

*rodrigosubodani@gmail.com

The car insurance claim process is usually done physically or manually. By automating, it can reduce the time taken and pay the correct amount while reducing claim leakage. The model proposed by this study initially identifies whether the concerned car part is damaged or not (Task 1). Then the damaged location is identified as Front, Rear or Side (Task 2) and then checks whether the part has to be repaired or replaced (Task 3). Finally, the model estimates the cost of the damage (Task 4) based on the external view. As compared with preceding studies, we improved the task 1 and 2 model performance by tuning hyperparameters and extended the research scope by assessing the damage severity classifying as repaired or replaced. The preprocessed dataset is available under the open data commons attribution license and the images are rearranged for task 3 and hand-labeled with a price tag for all the images. Scratched convolutional neural network and transfer learning models (VGG16, Xception, InceptionV3) were used for image classification. Transfer learning VGG16 model resulted in the highest accuracy. Testing accuracy for tasks 1,2 and 3 are 93%,75% and 76%, respectively. To estimate the cost instead of using metadata models, a vision-based regression model was proposed that extracted image features using a pre-trained model and passed them through a fine-tuned regression model. Two approaches were proposed for task 4, fitting one regression model after performing task 3, and two separate models for repair and replace. Rather than implementing one regression model, the implementation of two regression models separately resulted in higher accuracy. Gradient boosting regression is selected as the best model when estimating price for the repair (MSE - 0.88) and XGboost regression for replacement (MSE – 0.80). This approach is applicable for damages like scratches and dents and not considered when the image has multiple damages.

Keywords: *Deep learning, Computer vision, Convolutional Neural Network, Vision based Regression, Damage assessment*

Paddy Crop Quality Monitoring and Mapping Application using Object Detection Techniques

W D Nilakshi Sandeepanie[1]*, Samadhi Rathnayake[1], Amali Gunasinghe[1]

[1]Department of Computing, Sri Lanka Institute of Information Technology, Yakkala, Sri Lanka

*sandeepanien@gmail.com

This research addresses the pressing challenge of crop diseases, a pervasive threat to global agricultural productivity and food security. The focus of the study is on the timely identification and management of diseases in paddy crops, employing the Osmo V3 device for image collection and the YOLO v8 (You Only Look Once Version 8) algorithm for automated disease detection. The primary objective is to develop a precise, efficient, and scalable solution to empower farmers with early disease detection capabilities for effective crop management. The comprehensive data collection process involved an assembly of approximately 5000 high-resolution images of paddy crops, sourced randomly from fields in the Western and North-Central provinces. The use of the DJI Osmo V3 device and a smart mobile phone ensured a diverse and representative dataset, laying a solid foundation for empirical analysis. The dataset labelling process employed bounding boxes to mark disease-affected areas, enhancing the specificity of the model. Subsequent preprocessing and augmentation techniques, including noise reduction and data augmentation, elevated the dataset's quality and robustness. The YOLO v8 model was chosen for its state-of-the-art performance and real-time processing capabilities. The model was trained for 20 epochs, showcasing its adaptability to various data sources, including multi-scaled images, mp4 data, and streaming data. Performance evaluation metrics, such as precision, recall, and Mean Average Precision (mAP), provided quantitative insights into the model's accuracy. Auto-generated graphs depicted the distribution of diseases among the dataset, while the convergence of Train Box Loss and Validation/Box Loss indicated the model's robust generalization. The developed model not only demonstrated promising results in disease detection, including subtle symptoms, but also exhibited adaptability to diverse data sources. Future work aims to optimize parameters, integrate remote sensing techniques, and expand the applicability to a broader range of diseases. In conclusion, this research significantly contributes to crop disease management and global food security by providing farmers with a sophisticated tool for timely and accurate disease detection.

Keywords: *machine learning, object detection, web development, YOLO v8, diseases, paddy cultivation*

A Machine Learning Approach for Predicting the Depression Status Among Undergraduates

D S Sonnadara[1]*, S D Viswakula[1], M D T Attygalle[1]

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

* dilshan.sonnadara13@gmail.com

Depression is a common mental health disorder that affects a significant portion of the global population. In Sri Lanka, depression is a growing concern, especially among undergraduates. In this study, a machine learning model was developed to predict the depression status of undergraduates. Data were collected from a sample of 363 undergraduates in the Faculty of Science, University of Colombo via a google form which included the Patient Health Questionnaire-9 (PHQ-9) and several questions pertaining to their background. It was found that 29.7% of the sample screened positive for depression. The collected data were then used to build and evaluate 13 different machine learning models. The results showed that the Gradient Boosting Classifier with an accuracy of 79% on the test set had the best accuracy as well as the best precision in predicting the depression status. Furthermore, feature importance analysis identified the overall life satisfaction and level of stress associated with academic activities as the most important predictors of depression. A novel approach was used to incorporate post stratification weights in building a machine learning model on a representative sample. The model has two main limitations: it can only predict two classes (whether a student is at risk of developing depression or not), and it was validated only for the students in the Faculty of Science of the University of Colombo. Future research can explore modeling for more than 2 classes of depression and expanding the model to other faculties and universities. The developed model can be used as a screening tool to identify the students who may be at risk of developing depression and can aid in providing targeted interventions to improve the mental health and well-being of undergraduates.

Keywords: *Depression, PHQ-9, Gradient Boosting Classifier, Machine Learning*

Predicting the Outcome of T20 International Matches Based on Historical Statistics and Crowd Opinions

S P P M Sudasinghe[1]*, S D Viswakula[1], G P Lakraj[1]

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

*pavanthisppm@gmail.com

This study addresses the challenge of predicting the outcome of T20 international cricket matches prior to their commencement by proposing a machine learning approach. Twitter, which proves its capacity to produce real-time updates, was integrated with historical data for this purpose. Historical data and match-related tweets were scraped from 2011-01-01 to 2022-10-14, and appropriate features were derived from those scraped tweets. In order to derive one of those features, the study needed a sentiment analysis. The sentiment analysis results showed that the fine-tuned RoBERTa-based model outclassed the models created using LSTM (F1 Score: 87.8%) and VADER (F1 Score: 93.1%) with an F1 Score of 95.3%. After adding those derived variables, three datasets were formed: one with historical data, another with Twitter data, and a third combining both. Then multiple machine learning algorithms (Logistic Regression, SVM, Naive Bayes, KNN, Random Forest, and XGBoost) were trained and evaluated on these datasets. For all three datasets, the XGBoost algorithm emerged as the best classifier. Eliminating highly correlated and less important variables enhanced the models' performances. Furthermore, the findings of the study highlighted the predictive power of data gathered from the Twitter platform. The XGBoost Twitter data model (F1 Score: 71.5%) outperformed the XGBoost historical data model (F1 Score: 65.5%). Moreover, the XGBoost model developed using both historical data and Twitter data surpassed individual data models with a 73.7% F1 score. Impressively, this best model performed better than bookmakers' predictions for the T20I World Cup 2022 by accurately predicting the winners of 11 out of 14 matches. The bookmakers were able to correctly predict only nine matches.

Keywords: *Cricket, Sports Statistics, Sentiment Analysis, Natural Language Processing, Machine Learning*

The Study on Malaria Cases Diagnosed in Sri Lanka after Returning from Malaria-Endemic Countries

B D S Suwaris[1*], J H D S P Tissera[1], Pubudu Chulasiri[2]

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

[2]Anti Malaria Campaign, Ministry of Health, Colombo 05, Sri Lanka

*danushka.san96@gmail.com

Due to the presence of imported malaria cases in Sri Lanka, maintaining the malaria free status is at risk as it is compelling to spend more money and time on the disease. Thus, reliable screening and diagnosis methods are necessary apart from the well-known expensive tests to reduce the involved cost whilst management of severe cases has also become important. Thus, the main objective of this study is to identify significantly associated factors of the disease, and then to develop an accurate classifier to recognize the disease prone malaria species and the severity condition among the people who are returning from malaria endemic countries using the advanced machine learning techniques. Extreme Gradient Boosting using Random Forest classifier as the base learner with identified significant risk factors performed well with the highest accuracy of 80.43% and reasonable sensitivities. Also, Random Forest classifier fitted on synthesized data with synthetic minority oversampling technique is identified as the best method in detecting the severe infections with the accuracy of 95.65% along with a reasonable sensitivity of 67%. The findings with respect to the classifiers suggest that vulnerability towards the severity is high for malaria infected people who have been to African countries, while the species infect vary with the countries and occupations. The people who tested positive for plasmodium falciparum malaria infections show high vulnerability towards the severity, and the vulnerability towards severity increases with age while different groups show different levels of vulnerability for different species. Further, females show high vulnerability towards severe infections while species of infections vary across the gender. The patients who have a negative malaria history also show high vulnerability towards severe malaria infections. Consultation delays have been identified to have a significant impact in determining the species infect, while high diagnosing delays show high vulnerabilities towards severe infections.

Keywords: *accuracy, diagnosis, imported, malaria, Plasmodium falciparum, severity*

Convolutional Neural Network and Vision Transformer Models with Dynamic Attention for Emotion Recognition in Facial Dynamics-Based Stress Detection

Janudi Ranasinghe[1*], Amantha Jayathilake[1], Samadhi Rathnayake[2], Devanshi Ganegoda[2]

[1]*Department of Computer Systems and Software Engineering, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka*

[2]*Department of Information Technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka*

*rjanudi@gmail.com

This research paper explores stress detection through facial dynamics, employing Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Notably, we implemented a dynamic hierarchical attention mechanism in both models to enhance accuracy in capturing crucial facial features while mitigating noise. In today's dynamic world, where the impact of stress on well-being and productivity has gained paramount importance, this research introduces a stress detection and management application grounded in facial dynamics analysis. Facial expressions serve as potent indicators of emotions, including stress, allowing for non-intrusive real-time assessments. The study addresses the challenge of directing attention to relevant facial features through a dynamic hierarchical attention mechanism. Our novel approach incorporates dynamic attention weights for stress detection, assigning varying importance to facial features based on their contribution to stress levels, thereby enhancing model robustness and generalizability. Contributions include utilizing facial dynamics for stress detection, introducing dynamic attention mechanisms, and investigating their effects. Leveraging CNNs and ViTs architectures, the ViT model's global relationship capture capability complements CNN's feature extraction. Evaluation on the Fer2013 dataset, spanning diverse facial expressions, reveals significant accuracy improvements, with the CNN model achieving 0.70 accuracy and the ViT model outperforming at 0.8012, both after data augmentation. The real-time applicability of the models is showcased through webcam-based emotion detection, bridging the gap between research and practical deployment. Emotion categories are accurately converted into stress levels, guided by psychological expertise. Future implications extend beyond stress monitoring to encompass emotion recognition, mental health assessment, and enhanced human-computer interaction. By effectively addressing the challenge of directing attention to pertinent facial features and harnessing the potential of dynamic attention mechanisms, this research makes a substantial contribution to robust stress detection systems. Furthermore, its insights pave the way for broader applications in effective computing, enriching real-time emotional analysis.

Keywords: *Attention Mechanisms, Convolutional Neural Network (CNN), Visual Transformers (ViTs)*

Comparative Analysis of Machine Learning Algorithms for Breast Cancer Tumor Grade Prediction

M D T U Ranasinghe[1]*, J H D S P Tissera[1]

[1]Department of Statistics, University of Colombo, Colombo, Sri Lanka
*tharushi.u.ranasinghe2@gmail.com

The escalating incidence of breast cancer in Sri Lanka over recent decades underscores its prominence as a pressing public health issue. The assessment of tumor grade involves capturing morphological characteristics signifying malignancy and aggressiveness, with grades ranging from I to IV, determined through examination of tissue samples, where Grade I represents less aggressive normal cells and Grade IV indicates the most abnormal, aggressive, and rapidly proliferating cells. This study endeavours to address the pressing need for improved prediction of breast cancer tumor grade through a comprehensive analysis. A range of classical statistical models, including the proportional odds model and multinomial logistic regression model, were evaluated for predicting tumor grade. Subsequently, various machine learning models namely Random Forest, Decision Tree, K-Nearest Neighbours, Support Vector Machine, and Gradient Boosting were assessed. In response to the inherent data imbalance, resampling techniques were utilized including random under-sampling, random oversampling, SMOTE, SMOTE-Tomek, SMOTE-ENN, and ADASYN to balance the dataset. Parameters were tuned using grid search with 5-fold cross-validation. An ordinal classifier was applied to account for the ordinal nature of the tumor grades. The evaluation criteria for model selection included the Mean Multi-class Area Under the Curve (MAUC), F1 score, precision, sensitivity, and accuracy. Ultimately, the Gradient Boosting model with SMOTE-ENN emerged as the best-performing model for predicting tumor grade. This model achieved an accuracy rate of 96%, an F1 score of 96%, and a MAUC value of 0.845. Notably significant factors contributing to the model's predictive performance included histology, age, district, religion, and marital status. In conclusion, this study presents a powerful machine learning model, Gradient Boosting with SMOTE-ENN, for precise prediction of breast cancer tumor grade in Sri Lanka. The findings hold promise for improving clinical decision-making in breast cancer management, thereby improving patient outcomes, and informing healthcare policies.

Keywords: *tumor grade, ordinal classifier, SMOTE, SMOTE-ENN, SMOTE-Tomek, ADASYN*

Suspicious Activity Likelihood Detection, Indication & Tracking, and Alert System to Prevent Crimes

E Sankeetha[1*], T S W Walagedara[1], A J O Geenath[1], G R S D Gunasinghe[1], D S Alwis[1]

[1]School of Computing IIT, Colombo 00400, Sri Lanka

*sankeetha.20210498@iit.ac.lk

Reported crimes have increased and estimated statistics show that there are 770 million surveillance cameras installed around the world as of 2019, and there will be one billion of them shortly. There is a pressing need to reduce crimes and to automate monitoring and detection methods with higher accuracy and speedier warnings. The project aims to prevent crimes by integrating theft, violence, intrusion, and weapon detection with the incorporation of person re-identification (Re-ID) - enabling the system to remotely identify individuals within a predetermined area while into a single automated system - estimating the likelihood of threat, indicating, and alerting. The models were developed to detect four types of suspicious activities using diverse datasets, including Hockey Fight, Movies Fight, SCVD, AIRTLab, Weapon Detection2, and UCF Crime. While the weapon and intrusion detection model were built using You Only Look Once (YOLO) architecture, violence, and robbery were built on top of the MobileNet model - a pre-trained base model. The output was fed to a model using transfer learning methods with Bidirectional Long Short-Term Memory (LSTM) and a sequence of Dropout and Dense layers. The incorporated person re-identification model uses YOLO to detect objects in each frame and Deep Simple Online and Realtime (DeepSORT) to track those objects. It was found that the neural networks and transfer learning techniques improved the accuracy over time, enabling the models to learn from prior data and detect suspicious activities. The highest accuracies achieved for robbery, violence, weapon, and intrusion detection are 78%, 91% 85%, and 83%, respectively. To minimize false alerts, a feedback mechanism will be implemented which would refine the algorithms, mitigating false positives and negatives, thereby elevating the performance. The comprehensive novel approach of integrating diverse technologies enhances safety and reduces the crime rate on a large scale by detecting and addressing potential threats.

Keywords: *LSTM, YOLO, Deep SORT, Re-ID, MobileNet*

Identifying the Frequently Occurring Aquatic Insect Assemblages Using Association Rule Mining Techniques

K N W Tennakoon[1*], M D T Attygalle[1], P K T N S Pallewatta[2], M S Kanakarathna[3]

[1]Department of Statistics, University of Colombo, Colombo, Sri Lanka
[2]Department of Zoology and Environment Sciences, University of Colombo, Colombo, Sri Lanka
[3]Sydney, NSW, 2000, Australia
*kasunineranjana321@gmail.com

Aquatic insect compositions within assemblages are complex and depend on habitats and weather conditions. An in-depth analysis of such compositions can unveil information that would help understand various habitats, ecosystems, and environmental aspects. Only a few studies have been done in Sri Lanka on aquatic insects, where the focus has been on descriptive-analytical methods. This research is based on association rule mining, relating to Market Basket Analysis (MBA), to determine the frequently occurring freshwater insect assemblages. The dataset consisted of 372 sets of insects, from 19 insect groups, obtained across 5 Land Use Patterns (LUP), 3 elevations, and 6 sampling visits during a year. Compared with a typical MBA, the 19 insect groups were the items, and 372 sets were the transactions of the market items considering the presence and absence of each species group. Apriori Algorithm is used with a minimum support value of 0.1 and a confidence value of 0.8. Support is defined as the most frequent species group set or association rule between species group sets. Confidence is defined as the expected likelihood between two species group sets under the association rule. Due to the relatively low sample size of 372, the bootstrap sampling technique was applied to obtain reliable results. The analysis revealed that the species group set 'Caddisflies' was the most frequently occurring species group set. The species group sets 'Mayflies and Caddisflies' were found to be the most commonly coexisting pair. Moreover, the two highest frequent association rules were, if a Caddisfly is present a Mayfly exists, and if a Mayfly is present a Caddiafly exists. Findings were independent of LUPs and the visits. This study illustrates the use of association rule mining techniques to uncover aquatic insect assemblages and their co-existence that can be used as bio-indicators sensitive to water quality.

Keywords: *Apriori Algorithm, Aquatic Insects, Assemblages, Association Rule Mining, Market Basket Analysis*

A Study on Mental Disorders Among Psychiatric Patients in Kalutara District: A Hospital-Based Study

R A C Tharundi[1]*, A A Sunethra[1]

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

*chathunitharundi909@gmail.com

The study addresses the common mental health conditions in contemporary society, seeking to identify key demographic, socio-economic, lifestyle, and health-related factors associated with common psychiatric disorders. Using a systematic sampling technique, data were collected through interviews with 208 patients visiting three main hospitals in Kalutara district over one month. The identified psychiatric disorders included mood, psychotic disorders, psychotic illness, behavioural disorders, and anxiety. The analysis involves descriptive and inferential methods for identifying the factors associated with the disorders. Further, machine learning and deep learning techniques were employed to confirm the results observed in the initial study. Employing a questionnaire for data collection and Python software for analysis, the study focuses on a nominal response variable classifying disorder types into five categories. The chi-squared test and G-test were applied to identify associated factors. Subsequently, a predictive analysis employing five multiclass-classification models (K Nearest Neighbour, Random Forest, Categorical Naïve Bayes, Support Vector Machine, and XGBoost) and a deep-learning neural network model revealed that the XGBoost model has the best performance with a recall value of 84%. The feature importance helped to increase the recall value of the XGBoost model. The oversampling technique SMOTE was used for handling the class imbalance of the response variable and the grid search hyper-parameter tuning technique was used on the trained data using k-fold cross-validation. The results confirmed the accuracy of the preliminary study, emphasizing the impact of demographic variables such as gender, working status, marital status, number of children, family history of diseases, and sleeping disabilities on the response variable of disorder type. This study underscores the comprehensive approach taken to uncover factors associated with psychiatric disorders and the efficacy of advanced modeling techniques in predictive analysis for mental health conditions.

Keywords: *psychiatric disorders, systematic sampling, Chi-squared test, G-test, predictive analysis, machine-learning, XGBoost.*

Using Causal Inference in Portfolio Optimization: A Case Study

S I Ubayawickrema[1]*, K M M H Siriwardana[2], S D Viswakula[1]

[1]Department of Statistics, University of Colombo, Colombo 00300, Sri Lanka

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

*imalsha.ubayawickrema@gmail.com

Portfolio optimization has become a prominent issue in the present-day financial world. It allows the investors to make informed decisions about which assets to include in their portfolios, how much to invest in each asset, and when to make changes to their investments. This study aims to showcase the importance of using causal inference in portfolio optimization, since it is frequently influenced by the changing market interventions. A Bayesian Structural Time Series (BSTS) model is used to measure the causal impact of COVID-19 pandemic on 10 United States (US) stocks during the time horizon starting from Jan 3, 2017, to Nov 23, 2022. The stocks are based on five sectors: energy, finance, real estate, technology, and utilities and were chosen to balance out their risks when combined. Together, these sectors account for about 50% of the Standard and Poor's 500 (S&P500), a stock market index that is based on the 500 largest companies in the US. Two companies were chosen as the portfolio's securities from each industry. It was identified that from the 5 sectors, technology and real estate are among the most benefited sectors by the pandemic which indicated a positive impact of 113.35% and 39.06% in the securities selected from the respective sectors. The least benefited is the financial sector while energy and utilities sectors did not get causally impacted significantly. Upon completion of the study, it was concluded that causal impact insights derived by predicting the counterfactual series of an intervention is an effective tool in making informative business decisions in portfolio optimization under changing market conditions.

Keywords: *Portfolio Optimization, Causal Inference, Bayesian Structural Time Series model*

Feature Engineering and Music Genre Classification: A Showdown between XGBoost and Neural Networks

W M H G T C K Weerakoon[1]*, M S H Peiris[1], P B S N Ariyathilake[2]

[1]Department of Computer Science, Faculty of Science, University of Ruhuna, Matara, Sri Lanka

[2]Department of Remote Sensing and GIS, Faculty of Geomatics, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka

*weerakoonwmh@dcs.ruh.ac.lk

Music is a universal language and holds profound cultural significance. This study aims to provide an understanding of performance differentiation by employing gradient boosting algorithms (XGBoost) and neural networks (NN) in music genre classification when using feature engineering. Further, this study signifies that feature engineering plays a major role in final classification accuracies. Hence, Music Information Retrieval (MIR) techniques were employed to capture data pertaining to 1000 mp3 music tracks, spanning diverse genres including Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, and Rock. The dataset boasts an array of features, comprising tempo, beats, chroma short-term Fourier transformations, RMSE (Root-Mean Squared Error), spectral centroid, spectral bandwidth, rolloff, zero-crossing rate, and MFCC (Mel-Frequency Cepstral Coefficients), extracted through the Librosa Python library. Feature engineering techniques, encompassing Correlation Analysis (CA) and Principal Component Analysis (PCA), were applied to enrich classification accuracy. Three experiments were carried out with constant parameters for both XGBoost and Neural Networks. Experiment 01 (Exp-01) was conducted with pre-processed data, whereas experiment 02 (Exp-02) and experiment 03 (Exp-03) were performed with CA and PCA respectively. Training and validation accuracy, precision, recall and F1-score were obtained as benchmarking metrics for each experiment. Training accuracies of the three experiments for both XGBoost and NN models were Exp-01:100% ,63%, Exp-02: 98%, 90% and Exp-03: 100%, 95% respectively. Validation accuracies were Exp-01:67%, 67%, Exp-02:99%, 90% and Exp-03:52%, 58%, respectively. When XGBoost and Neural Networks are compared with the influence of CA and PCA, the XGBoost classifier with CA as feature engineering produces promising results for music genre categorization. As a future prospect, the concluded model will be implemented, and an ablation study for Exp-02 will be conducted to further assess the neural network.

Keywords: *Feature Engineering, XGBoost, Neural Networks, Correlation Analysis, PCA*

Skin Cancer Image Classification using Vision Transformer Network

W M D D B Wijesundara[1]*, C T Wannige[1], M K S Madushika[2], P L A N Liyanage[3]

[1]Department of Computer Science, Faculty of Medicine, University of Ruhuna, Matara, Sri Lanka

[2]Revolution Aerospace, Brisbane, Australia

[3]Department of Community Medicine, Faculty of Medicine, University of Ruhuna, Matara, Srilanka

*bandarad@usci.ruh.ac.lk

Skin cancer is defined by abnormal skin cell proliferation and, if ignored or undiagnosed, can lead to significant impacts on an individual's life, including disfigurement, impaired quality of life, and even death. Melanomas, as a concern with a poor prognosis, have shown increasing incidence in the recent past. In order to improve patient care and lessen the impact of the condition, early identification is essential. Several systematic screening programmes are designed to detect skin cancer in its earliest stages for a better outcome. In addition, many deep learning and machine learning techniques have been used to identify skin cancer via image processing. This study offers a thorough investigation of skin cancer categorization utilizing the Vision Transformer (ViT) to develop a high-accuracy model for early skin cancer identification. Generally, CNN-based models struggle with accurate skin lesions classification due to inconsistent lesion shape and loss of local feature attributes. As a solution, transformer-based models can be used as they exploit local and global characteristics, self-attention processes, and expressive long-range representations. The study uses the HAM10000 dataset, and to address dataset imbalances, the Synthetic Minority Over-sampling Technique (SMOTE) is employed. Data preprocessing includes image resizing and pixel value normalization. Augmentation techniques enhance model generalization. The training process employs a train-validation-test split (95%, 2.5%, and 2.5%) and an AdamW optimizer with 61 epochs. Model performance is evaluated using a training accuracy of 98.73%, validation accuracy of 98.23%, and a normalized confusion matrix. Precision, recall, and F1-score values provide insights into class-specific performance. Overall, the study underscores the potential of deep learning models for skin cancer classification, with room for further enhancements through strategic adjustments. Future work involves refining generalizations to new data, exploring new ViT architectures, and avoiding overfitting risks. With careful refinement, this deep learning model holds promise for enhancing image classification in practical applications and decision-making processes.

Keywords: *Skin cancer, Vision Transformer, Deep learning, Image classification, Dermatology.*

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.

9 773030 716005

ICDS 2023

Data Science in the Age of Artificial Intelligence