Lecture Notes on Data Engineering and Communications Technologies 114

Mukesh Saraswat · Harish Sharma · K. Balachandran · Joong Hoon Kim · Jagdish Chand Bansal *Editors*

Congress on Intelligent Systems

Proceedings of CIS 2021, Volume 1

Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 114

Series Editor

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It will publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

Indexed by SCOPUS, INSPEC, EI Compendex.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at https://link.springer.com/bookseries/15362

Mukesh Saraswat · Harish Sharma ·
K. Balachandran · Joong Hoon Kim ·
Jagdish Chand Bansal
Editors

Congress on Intelligent Systems

Proceedings of CIS 2021, Volume 1

Springer

Editors
Mukesh Saraswat
Department of Computer Science & Engineering and Information Technology
Jaypee Institute of Information Technology
Noida, India

Harish Sharma
Department of Computer Science and Engineering
Rajasthan Technical University
Kota, India

K. Balachandran
Department of Computer Science and Engineering
CHRIST (Deemed to be University)
Bengaluru, India

Joong Hoon Kim
Korea University
Seoul, Korea (Republic of)

Jagdish Chand Bansal
South Asian University
New Delhi, India

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-981-16-9415-8 ISBN 978-981-16-9416-5 (eBook)
https://doi.org/10.1007/978-981-16-9416-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This proceedings contains the papers presented at the 2nd Congress on Intelligent Systems (CIS 2021), organized by CHRIST (Deemed to be University), Bengaluru, India, and Soft Computing Research Society during September 4–5, 2021. Congress on Intelligent Systems (CIS 2021) invited ideas, developments, applications, experiences, and evaluations in intelligent systems from academicians, research scholars, and scientists. The conference deliberation included topics specified within its scope. The conference offered a platform for bringing forward extensive research and literature across the arena of intelligent systems. It provided an overview of the upcoming technologies. CIS 2021 provided a platform for leading experts to share their perceptions, provide supervision, and address participants' interrogations and concerns. CIS 2021 received 370 research submissions from 35 different countries, viz., Algeria, Bangladesh, Burkina Faso, China, Egypt, Ethiopia, Finland, India, Indonesia, Iran, Iraq, Kenya, Korea, The Democratic People's Republic of Madagascar, Malaysia, Mauritius, Mexico, Morocco, Nigeria, Peru, Romania, Russia, Serbia, Slovakia, South Africa, Spain, Switzerland, Ukraine, United Arab Emirates, UK, USA, Uzbekistan, Vietnam. The papers included topics about advanced areas in technology, artificial intelligence, machine learning, and data science. After a rigorous peer review with the help of program committee members and more than 100 external reviewers, 135 papers were approved.

CIS 2021 is a flagship event of the Soft Computing Research Society, India. The conference was inaugurated by Fr. Dr. Abraham VM, Honorable Vice-Chancellor, CHRIST (Deemed to be University), Bangalore, and Chief Patron, CIS 2021. Other eminent dignitaries include Prof. Joong Hoon Kim, General Chair, CIS 2021; Fr. Joseph Varghese, Patron, CIS 2021; and Prof. K. Balachandran, General Chair, CIS 2021. The conference witnessed keynote addresses from eminent speakers, namely Prof. Xin-She Yang, Middlesex University, The Burroughs, Hendon, London; Prof. P. Nagabhushan, Indian Institute of Information Technology Allahabad; Prof. J. C. Bansal, South Asian University, New Delhi, India; Prof. Lipo Wang, Nanyang Technological University, Singapore; and Prof. Nishchal K. Verma, Indian Institute of

Technology Kanpur, India. The organizers wish to thank Mr. Aninda Bose, Senior Editor, Springer Nature, New Delhi, India, for the support and guidance.

Noida, India	Mukesh Saraswat
Kota, India	Harish Sharma
Bengaluru, India	K. Balachandran
Seoul, Korea (Republic of)	Joong Hoon Kim
New Delhi, India	Jagdish Chand Bansal

Contents

The Extraction of Automated Vehicles Traffic Accident Factors and Scenarios Using Real-World Data 1
MinHee Kang, Jaein Song, and Keeyeon Hwang

Leaf Disease Identification in Rice Plants Using CNN Model 17
Allam Sushanth Reddy and Jyothi Thomas

Twitter Sentiment Analysis Based on Neural Network Techniques 33
Ashutosh Singal and Michael Moses Thiruthuvanathan

Enhanced Stock Market Prediction Using Hybrid LSTM Ensemble 49
Reuben Philip Roy and Michael Moses Thiruthuvanathan

Classifying Microarray Gene Expression Cancer Data Using Statistical Feature Selection and Machine Learning Methods 63
S. Alagukumar and T. Kathirvalavakumar

Pythagorean Fuzzy Information Measure with Application to Multicriteria Decision Making 79
Anjali Munde

Developing an Improved Software Architecture Framework for Smart Manufacturing 87
Gareth A. Gericke, Rangith B. Kuriakose, and Herman J. Vermaak

Intelligent Water Drops Algorithm Hand Calculation Using a Mathematical Function 101
Sathish Kumar Ravichandran, Archana Sasi, and Ramesh Vatambeti

A Study of Decision Tree Classifier to Predict Learner's Progression . . . 113
Savita Mohurle and Richa Pandey

An Overview of Blockchain and IoT in e-Healthcare System 123
S. V. Vandana AryaSomayajula and Ankur Goyal

Energy-Efficient ACO-DA Routing Protocol Based on IoEABC-PSO Clustering in WSN 139
M. Vasim Babu, C. N. S. Vinoth Kumar, B. Baranidharan, M. Madhusudhan Reddy, and R. Ramasamy

Modelling Critical Success Factors for Smart Grid Development in India 157
Archana, Ravi Shankar, and Shveta Singh

Stability Analysis of Emerged Seaside Perforated Quarter Circle Breakwater Using Soft Computing Techniques 177
Sreelakshmy Madhusoodhanan and Subba Rao

A Risk-Budgeted Portfolio Selection Strategy Using Novel Metaheuristic Optimization Approach 193
Mohammad Shahid, Zubair Ashraf, Mohd Shamim, Mohd Shamim Ansari, and Faisal Ahmad

An Optimization Reconfiguration Reactive Power Distribution Network Based on Improved Bat Algorithm 205
Thi-Kien Dao, Trinh-Dong Nguyen, Trong-The Nguyen, and Jothiswaran Thandapani

Analyzing a Raga-Based Bollywood Song: A Statistical Approach 217
Lopamudra Dutta and Soubhik Chakraborty

Security Prioritized Heterogeneous Earliest Finish Time Workflow Allocation Algorithm for Cloud Computing 233
Mahfooz Alam, Mohammad Shahid, and Suhel Mustajab

Dropout-VGG Based Convolutional Neural Network for Traffic Sign Categorization 247
Inderpreet Singh, Sunil Kr. Singh, Sudhakar Kumar, and Kriti Aggarwal

Internet-Based Healthcare Things Driven Deep Learning Algorithm for Detection and Classification of Cervical Cells 263
Shruti Suhas Kute, Amit Kumar Tyagi, Shaveta Malik, and Atharva Deshmukh

Load Balancing Algorithms in Cloud Computing Environment—An Effective Survey 279
N. Priya and S. Shanmuga Priya

Assessment of the Spatial Variability of Air Pollutant Concentrations at Industrial Background Stations in Malaysia Using Self-organizing Map (SOM) 291
Loong Chuen Lee and Hukil Sino

An Approach for Enhancing Security of Data over Cloud Using Multilevel Algorithm 305
Binita Thakkar and Blessy Thankachan

Advanced Spam Detection Using NLP and Deep Learning 319
Aditya Anil, Ananya Sajwan, Lalitha Ramchandar, and N. Subhashini

A Systematic Literature Review on Image Preprocessing and Feature Extraction Techniques in Precision Agriculture 333
G. Sharmila and Kavitha Rajamohan

A Comprehensive Study on Computer-Aided Cataract Detection, Classification, and Management Using Artificial Intelligence 355
Binju Saju and R. Rajesh

Improving Black Hole Algorithm Performance by Coupling with Genetic Algorithm for Feature Selection 367
Hrushikesh Bhosale, Prasad Ovhal, Aamod Sane, and Jayaraman K. Valadi

Brain Tumor Analysis and Reconstruction Using Machine Learning 381
Priyanka Sharma, Dinesh Goyal, and Neeraj Tiwari

Ordered Ensemble Classifier Chain for Image and Emotion Classification 395
Puneet Himthani, Puneet Gurbani, Kapil Dev Raghuwanshi, Gopal Patidar, and Nitin Kumar Mishra

A Novel Deep Learning SFR Model for FR-SSPP at Varied Capturing Conditions and Illumination Invariant 407
R. Bhuvaneshwari, P. Geetha, M. S. Karthika Devi, S. Karthik, G. A. Shravan, and J. Surenthernath

Attention-Based Ensemble Deep Learning Technique for Prediction of Sea Surface Temperature 425
Ashapurna Marndi and G. K. Patra

Political Optimizer-Based Optimal Integration of Soft Open Points and Renewable Sources for Improving Resilience in Radial Distribution System 439
D. Sreenivasulu Reddy and Varaprasad Janamala

Face and Emotion Recognition from Real-Time Facial Expressions Using Deep Learning Algorithms 451
Shrinitha Monica and R. Roseline Mary

A Real-Time Traffic Jam Detection and Notification System Using Deep Learning Convolutional Networks 461
Sedish Seegolam and Sameerchand Pudaruth

Design of a Robotic Flexible Actuator Based on Layer Jamming 477
Kristian Kowalski and Emanuele Lindo Secco

Sentiment Analysis on Diabetes Diagnosis Health Care Using Machine Learning Technique 491
P. Nagaraj, P. Deepalakshmi, V. Muneeswaran, and K. Muthamil Sudar

Predicting the Health of the System Based on the Sounds 503
Manisha Pai and Annapurna P. Patil

A Model Based on Convolutional Neural Network (CNN) for Vehicle Classification 519
F. M. Javed Mehedi Shamrat, Sovon Chakraborty, Saima Afrin, Md. Shakil Moharram, Mahdia Amina, and Tonmoy Roy

A Transfer Learning Approach for Face Recognition Using Average Pooling and MobileNetV2 531
F. M. Javed Mehedi Shamrat, Sovon Chakraborty, Md. Shakil Moharram, Tonmoy Roy, Masudur Rahman, and Biraj Saha Aronya

A Deep Learning Approach for Splicing Detection in Digital Audios 543
Akanksha Chuchra, Mandeep Kaur, and Savita Gupta

Multi-criteria Decision Theory-Based Cyber Foraging Peer Selection for Content Streaming 559
Parisa Tabassum, Abdullah Umar Nasib, and Md. Golam Rabiul Alam

Visualizing Missing Data: COVID-2019 573
K. Lavanya, G. Raja Gopal, M. Bhargavi, and V. Akhil

Study of Impact of COVID-19 on Students Education 585
Deepali A. Mahajan and C. Namrata Mahender

A Framework for Analyzing Crime Dataset in *R* Using Unsupervised Optimized *K*-means Clustering Technique 593
K. Vignesh, P. Nagaraj, V. Muneeswaran, S. Selva Birunda, S. Ishwarya Lakshmi, and R. Aishwarya

Seed Set Selection in Social Networks Using Community Detection and Neighbourhood Distinctness 609
Sanjeev Sharma and Sanjay Kumar

An Optimized Active Learning TCM-KNN Algorithm Based on Intrusion Detection System 621
Reenu Batra, Manish Mahajan, and Amit Goel

Ensemble Model of Machine Learning for Integrating Risk in Software Effort Estimation 635
Ramakrishnan Natarajan and K. Balachandran

A Questionnaire-Based Analysis of Network Forensic Tools 645
Rachana Yogesh Patil and Manjiri Ranjanikar

On the Industrial Clustering: A View from an Agent-Based Version of Krugman Model 653
Smicha Ait Amokhtar and Nadjia El Saadi

An Efficient Comparison on Machine Learning and Deep Neural Networks in Epileptic Seizure Prediction 677
R. Roseline Mary, B. S. E. Zoraida, and B. Ramamurthy

Analysis of Remote Sensing Satellite Imagery for Crop Yield Mapping Using Machine Learning Techniques 689
M. Sarith Divakar, M. Sudheep Elayidom, and R. Rajesh

Analysis of Lung Cancer Prediction at an Early Stage: A Systematic Review 701
Shweta Agarwal and Chander Prabha

Framework for Estimating Software Cost Using Improved Machine Learning Approach 713
Sangeetha Govinda

Deducing Water Quality Index (WQI) by Comparative Supervised Machine Learning Regression Techniques for India Region 727
Sujatha Arun Kokatnoor, Vandana Reddy, and K. Balachandran

Artificial Ecosystem-Based Optimization for Optimal Location and Sizing of Solar Photovoltaic Distribution Generation in Agriculture Feeders 743
U. Kamal Kumar and Varaprasad Janamala

Optimized Segmentation Technique for Detecting PCOS in Ultrasound Images 759
S. Jeevitha and N. Priya

Implementation of Morphological Gradient Algorithm for Edge Detection 773
Mirupala Aarthi Vardhan Rao, Debasish Mukherjee, and S. Savitha

Support Vector Machine Performance Improvements by Using Sine Cosine Algorithm 791
Miodrag Zivkovic, Nikola Vukobrat, Amit Chhabra, Tarik A. Rashid, K. Venkatachalam, and Nebojsa Bacanin

French COVID-19 Tweets Classification Using FlauBERT Layers 805
Sadouanouan Malo, Thierry Roger Bayala, and Zakaria Kinda

A Novel Weighted Extreme Learning Machine for Highly Imbalanced Multiclass Classification 817
Siddhant Baldota and Deepti Aggarwal

An Enhanced Pixel Intensity Range-Based Reversible Data Hiding Scheme for Interpolated Images 831
Rama Singh and Ankita Vaish

UAV Collaboration for Autonomous Target Capture 847
Lima Agnel Tony, Shuvrangshu Jana, V. P. Varun, Shantam Shorewala, B. V. Vidyadhara, Mohitvishnu S. Gadde, Abhishek Kashyap, Rahul Ravichandran, Raghu Krishnapuram, and Debasish Ghose

A Leaf Image-Based Automated Disease Detection Model 863
Aditi Ghosh and Parthajit Roy

Multi-agent Cooperative Framework for Autonomous Wall Construction 877
Kumar Ankit, Lima Agnel Tony, Shuvrangshu Jana, and Debasish Ghose

Construction of a Convex Polyhedron from a Lemniscatic Torus 895
Ricardo Velezmoro-León, Robert Ipanaqué-Chero, Felícita M. Velásquez-Fernández, and Jorge Jimenez Gomez

An Ant System Algorithm Based on Dynamic Pheromone Evaporation Rate for Solving 0/1 Knapsack Problem 911
Ruchi Chauhan, Nirmala Sharma, and Harish Sharma

Ant System Algorithm with Output Validation for Solving 0/1 Knapsack Problem 921
Ruchi Chauhan, Nirmala Sharma, and Harish Sharma

Fitness-Based PSO for Large-Scale Job-Shop Scheduling Problem 933
Kavita Sharma and P. C. Gupta

Author Index 949

About the Editors

Dr. Mukesh Saraswat is an associate professor at Jaypee Institute of Information Technology, Noida, India. Dr. Saraswat has obtained his Ph.D. in Computer Science and Engineering from ABV-IIITM Gwalior, India. He has more than 18 years of teaching and research experience. He has guided two Ph.D. students, more than 50 M.Tech. and B.Tech. dissertations, and presently guiding four Ph.D. students. He has published more than 50 journal and conference papers in the area of image processing, pattern recognition, data mining, and soft computing. He was the part of successfully completed DRDE funded project on image analysis and SERB-DST (New Delhi) funded project on Histopathological Image Analysis. He is currently running a projects funded by and Collaborative Research Scheme (CRS), Under TEQIP III (RTU-ATU) on Smile. He has been an active member of many organizing committees of various conferences and workshops. He is also a guest editor of the *Array*, *Journal of Swarm Intelligence*, and *Journal of Intelligent Engineering Informatics*. He is an active member of IEEE, ACM, and CSI Professional Bodies. His research areas include image processing, pattern recognition, mining, and soft computing.

Harish Sharma is an associate professor at Rajasthan Technical University, Kota, in the Department of Computer Science and Engineering. He has worked at Vardhaman Mahaveer Open University, Kota, and Government Engineering College, Jhalawar. He received his B.Tech. and M.Tech degree in Computer Engg. from Government Engineering College, Kota, and Rajasthan Technical University, Kota, in 2003 and 2009, respectively. He obtained his Ph.D. from ABV—Indian Institute of Information Technology and Management, Gwalior, India. He is secretary and one of the founder member of Soft Computing Research Society of India. He is a life time member of Cryptology Research Society of India, ISI, Kolkata. He is an associate editor of "International Journal of Swarm Intelligence (IJSI)" published by Inderscience. He has also edited special issues of the many reputed journals like *Memetic Computing*, *Journal of Experimental and Theoretical Artificial Intelligence*, *Evolutionary Intelligence*, etc. His primary area of interest is nature-inspired optimization techniques. He has contributed in more than 105 papers published in various international journals and conferences.

Dr. K. Balachandran is currently a professor and the head CSE at CHRIST (Deemed to be University), Bengaluru, India. He has total 38 years' experience in Research, Academia, and Industry. He served as a senior scientific officer in the Research and Development unit of Department of Atomic Energy for 20 years. His research interest includes data mining, artificial neural networks, soft computing, and artificial intelligence. He has published more than fifty articles in well-known SCI/SCOPUS indexed international journals and conferences and attended several national and international conferences and workshops. He has authored/edited four books in the area of computer science.

Prof. Joong Hoon Kim a faculty of Korea University in the School of Civil, Environmental, and Architectural Engineering, obtained his Ph.D. degree from the University of Texas at Austin in 1992 with the thesis title "Optimal replacement/rehabilitation model for water distribution systems." Professor Kim's major areas of interest include: optimal design and management of water distribution systems, application of optimization techniques to various engineering problems, and development and application of evolutionary algorithms. His publication includes *A New Heuristic Optimization Algorithm: Harmony Search*, Simulation, February 2001, Vol. 76, pp 60–68, which has been cited over 2300 times by other journals of diverse research areas. His keynote speeches include "Optimization Algorithms as Tools for Hydrological Science" in the Annual Meeting of Asia Oceania Geosciences Society held in Brisbane, Australia, in June of 2013, "Recent Advances in Harmony Search Algorithm" in the 4th Global Congress on Intelligent Systems (GCIS 2013) held in Hong Kong, China, in December of 2013, and "Improving the convergence of Harmony Search Algorithm and its variants" in the 4th International Conference on Soft Computing For Problem Solving (SOCPROS 2014) held in Silchar, India in December of 2014. He hosted the 6th Conference of the Asia Pacific Association of Hydrology and Water Resources (APHW 2013) and the 2nd International Conference on Harmony Search Algorithm (ICHSA 2015). And, he is hosting another one, the 12th International Conference on Hydroinformatics (HIC 2016) in August of 2016.

Dr. Jagdish Chand Bansal is an associate professor at South Asian University, New Delhi, and a visiting faculty at Maths and Computer Science, Liverpool Hope University, UK. Dr. Bansal has obtained his Ph.D. in Mathematics from IIT Roorkee. Before joining SAU, New Delhi, he has worked as an assistant professor at ABV—Indian Institute of Information Technology and Management Gwalior and BITS Pilani. He is the series editor of the book series *Algorithms for Intelligent Systems* (AIS) published by Springer. He is the editor in chief of *International Journal of Swarm Intelligence* (IJSI) published by Inderscience. He is also the associate editor of *IEEE Access* published by IEEE. He is the steering committee member and the general chair of the annual conference series SocProS. He is the general secretary of Soft

Computing Research Society (SCRS). His primary area of interest is swarm intelligence and nature-inspired optimization techniques. Recently, he proposed a fission–fusion social structure-based optimization algorithm, Spider Monkey Optimization (SMO), which is being applied to various problems from engineering domain. He has published more than 70 research papers in various international journals/conferences. He has supervised Ph.D. theses from ABV-IIITM Gwalior and SAU, New Delhi. He has also received Gold Medal at UG and PG levels.

The Extraction of Automated Vehicles Traffic Accident Factors and Scenarios Using Real-World Data

MinHee Kang, Jaein Song, and Keeyeon Hwang

Abstract As automated vehicles (AVs) approach commercialization, the fact that the SAFETY problem becomes more concentrated is not controversial. Depending on this issue, the scenarios research that can ensure safety and are related to vehicle safety assessments are essential. In this paper, based on 'report of traffic collision involving an AVs' provided by California DMV (Department of Motor Vehicles), we extract the major factors for identifying AVs traffic accidents to derive basic AVs traffic accident scenarios by employing the random forest, one of the machine learning. As a result, we have found the importance of the pre-collision movement of neighboring vehicles to AVs and inferred that they are related to collision time (TTC). Based on these factors, we derived scenarios and confirm that AVs rear-end collisions of neighboring vehicles usually occur when AVs are ahead in passing, changing lanes, and merge situations. While most accident determinants and scenarios are expected to be similar to those between human driving vehicles (HVs), AVs are expected to reduce accident rates because 'AVs do not cause accidents.'

Keywords Automated vehicle · Real datasets · Traffic accident · Road safety · Traffic scenario · Machine learning

1 Introduction

AVs stand out in IT firms, unlike the traditional automobile manufacturing industry companies [1]. Especially, Waymo LLC preemptively is developing and occupying AVs technology and is being chased by the existing automobile manufacturing

M. Kang (✉) · J. Song · K. Hwang
Hongik University, Seoul 04066, Republic of Korea
e-mail: speakingbee@hanmail.net

J. Song
e-mail: wodlsthd@nate.com

K. Hwang
e-mail: keith@hongik.ac.kr

M. Saraswat et al. (eds.), *Congress on Intelligent Systems*, Lecture Notes on Data Engineering and Communications Technologies 114,
https://doi.org/10.1007/978-981-16-9416-5_1

industry as a latecomer [2]. This is judged that the IT companies that deal with data first occupied the AVs technologies due to the software characteristics of the AVs. However, the SAFETY issue is spotlighted due to pedestrian fatal crashes in operating Uber's AVs in 2018 [3–5]. Also, by emerging the study that AVs have the probability of collision even in ideal conditions [6], commercialization of AVs has been delaying. Although AVs clearly will be to take root in our society, ultimately the safety issues will eventually draw more attention in the future [7]. According to the issues, the safety paradigm has been changing from a passive safety system in the past to an active safety system that can prevent various accidents in advance [8]. One method of the active safety system is the prevention of traffic accidents in advance by learning numerous situations [9, 10], even if countless traffic accidents occur in real traffic conditions [11]. Furthermore, the recently proposed method for AVs assessment is the scenarios with ensuring safety. According to Ulbrich et al. [12], a scenario is the temporal development between several scenes in a sequence of scenes, actions, and events, as well as goals and values, specified to characterize this temporal development. Namely, these scenario studies that directly involved in car safety assessment are essential [13–15] and ambiguity about scenarios should need to minimize [18]. Along with these features, diverse studies that use real data for extracting scenarios are being carried out [16–18]. This is more relevant with this new paradigm shift in test and validation [19]. For instance, Erdogan et al. [19] pointed out that proposed scenarios are mostly hand-driven or expert input for AVs safety assessment. The authors stated that a myriad of scenarios is expected to be presented, so it is alone would not be suitable. In order to supplement the limitation, they proposed multiple classification algorithms and effectively extracted scenarios. Also, Webb et al. [20] presented an approach and methodology for AVs safety scenario assessment to reduce accident fatalities through safety and trust driving. With regard to using real data, Schwall et al. [16] demonstrated the safety of AVs on urban roads based on actual Waymo driving data.

In this paper, we propose a 2-step scenario-building method to improve safety by alleviating ambiguity. This is as follows: (1) Extract the major factors for determining AVs collision, and (2) based on extracted factors, draw the expected AVs traffic accident scenarios. The aim is to utilize traffic accidents data of AVs and machine learning methodology for the extraction of AVs traffic accident scenarios. The contribution is that (1) scenarios can be derived by using AVs accident data containing direct accidents, (2) as not human driving vehicles (HVs) data but real AVs accident data, empower actually AVs safety, (3) identify main decision factors through machine learning, and (4) be able to relieve dependence on the knowledge of experts which was deemed a blind spot.

This paper is organized as follows: In Sect. 2, we present collecting and preprocessing AVs accident data. Section 3 summarizes the method of using machine learning and describes the process of extraction to major decision factors. Based on this, we derive fundamental AVs accident scenarios in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Data Collecting and Preprocessing

In this section, we collect and preprocess the AVs traffic accident data.

2.1 Collecting AVs Accident Data

Currently, more than 13 production companies, including Waymo, GM, and Drive AI, are participating in the demonstration of autonomous driving in California, the USA. However, as the information about these test runs is confidential to the manufacturer, the manufacturer does not disclose data containing the driving information on its own. Although there is a limit to the collection of operating information data, the Department of Motor Vehicles (DMV) in California is releasing data on the 'traffic accident' situation that occurred in AVs [21]. The report shows AVs manufacturers, accident times, vehicle driving conditions (move and stop), autonomous driving modes, road users (pedestrians, etc.), weather, lighting, road conditions, information that can be inferred from the accident, etc.

Accordingly, we collected data on AVs accidents released by the State of California. They have data from 2014, but the data they are currently disclosing is data from 2019 onward. So, we contacted DMV and shared data for 2014–2018. As a result of checking the data, about 300 data existed, and the report format was changed around April 2018.

2.2 Preprocessing AVs Accident Data

We collected the accident history information provided by DMV and dataized what it regards will affect the accident situation. As previously described, the format of the report changed from 2018, so there was a limit to dataizing all data. As a result, data has been utilized since April 2018, with accounting for a total of 219 cases.

The data determined by the critical variables was referred to in Kang et al. [10]. First of all, we extracted accidents that occurred in autonomous driving mode through 'Mode' variable. The data corrected consists of 'Time,' 'AV injury,' 'Weather,' 'Lighting,' 'Roadway surface (Rs),' 'Roadway conditions (Rc),' 'Vehicle 1,2-Movement Preceding Collision (MPC),' 'Vehicle 1,2-Type of Collision (TC),' and 'Other associated factors (Other)' as shown in Table 1. Among them, the extra variables such as MPC 1–1 and 2–1 were two marked values in raw data, which were determined to be variables that affect the accident situation by reviewing the 'Accident Details' section. Consequently, we finally obtained 138 data consisting of 15 variables.

Table 1 AVs traffic accidents of features

Feature	Description
Mode	The driving mode of AVs including autonomous and conventional mode
Time	The time when AVs accident occurred
AV injury	The degree of injury to AVs in the accident including Minor, Major, MOD, etc.
Weather	The weather at the time of the accident
Lighting	The illumination of the surrounding environment at the time of the accident
Roadway surface (Rs)	The type of road surface at the time of the accident
Roadway condition (Rc)	The condition of the road surface at the time of the accident
Movement preceding collision (MPC) of vehicle 1	Vehicle 1 movement before the accident
Movement preceding collision (MPC) of vehicle 1–1	Vehicle 1 (additional) movement before the accident
Movement preceding collision (MPC) of vehicle 2	Vehicle 2 movement before the accident
Movement preceding collision (MPC) of vehicle 2–1	Vehicle 2 (additional) movement before the accident
Type of collision (TC) of vehicle 1	Type of collision that caused the accident
Type of collision (TC) of vehicle 1–1	
Type of collision (TC) of vehicle 2	
Type of collision (TC) of vehicle 2–1	
Other associated factors (other)	Other factors that may be associated with

3 Learning Method and Result

In this section, we extract important factors from AVs accidents by utilizing machine learning methodology for the purpose of the study.

3.1 Random Forest

Random forests [22] are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. In other words, while some trees of random forests have the probability of overfitting, it will be prevented by generating numerous trees.

Therefore, we utilize random forest technique that minimizes overfitting among the important variable extraction techniques to extract the main factors determining AVs traffic accidents. Among random forest importance variable extraction techniques, we deploy the 'Drop Columns Importance' method, which derives importance based on learning accuracy differences when learned except for certain variables.

3.2 Extraction of AVs Collision Factor Importance

Prior to applying the random forests, we discovered the following facts from the confirmation of refined data: (1) AVs did not crash with the object (e.g., pedestrian, etc.) in the autonomous driving mode, and (2) in a collision situation, the collision was caused by neighboring vehicles (HVs), not AVs. Therefore, we derived the feature importance of the variables by setting the dependent variable to 'AV injury' (see Table 2).

As a result of extraction (see Fig. 1), MPC2 and MPC1 were found to be the most important variables, with 24% and 22%, respectively. This is the movement of vehicles in pre-collision, and the movement of HVs (MPC1) is more important than AVs (MPC2). This is because AVs do not cause accidents, and most of the HVs around them are considered to cause accidents. Furthermore, the MPC itself is deemed to be due to the interaction between vehicles on the road. It is widely used as a vehicle safety indicator and is considered to have been derived as an important variable as it is related to time-to-collision (TTC) [23, 24], which means the time remaining immediately before the collision with the other vehicle. Next, 'Time' and 'Lighting' variables were ranked 3rd and 5th, which is deemed to be because the degree of injury in AVs operation varies depending on environmental conditions (day and night, lighting). This is related to be 'street's low-light at night' environmental conditions like Uber pedestrian fatal collision in 2018 [25]. Finally, the variables

Table 2 Learning environment of AVs traffic accidents

Feature	Class	Variable
AV injury	Output layer	0–4
Time	Input layer	0–3
Weather		0–7
Lighting		0–5
Roadway surface (Rs)		0–4
Roadway condition (Rc)		0–7
Movement preceding collision (MPC)		0–19
Type of collision (TC)		0–8
Other associated factors (other)		0–12

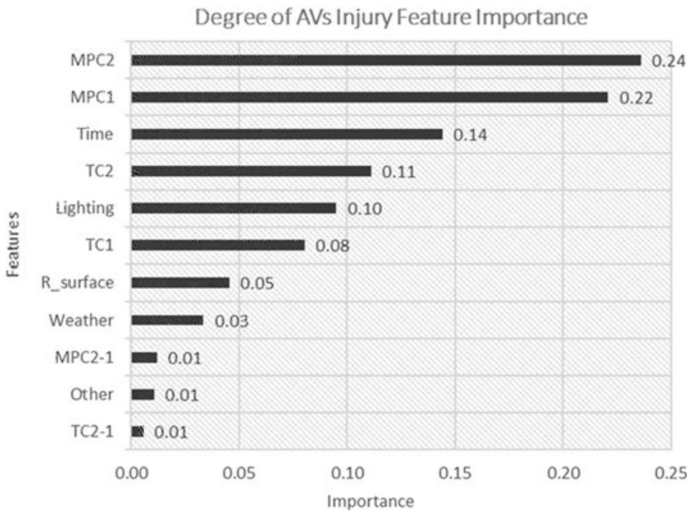


Fig. 1 Factor for determining AVs collision

for 'Type of Collision (TC)' ranked 4th and 6th, respectively. This is considered that the variables can have various effects depending on the seating position of the occupants in the vehicle. In this regard, studies of seating position in the HVs have been conducted [26–28], and many studies have carried out in the AVs sector too [29–33].

4 AVs Traffic Accident Scenarios

In this section, we derive a scenario for AVs accidents based on the factors previously extracted.

4.1 Scenario Combinations Based on Collision Situations

Scenarios are derived based on the most important factors, Movement Preceding Collision 1 and 2, as previously identified data. Among them, there will be a variety of movement preceding collisions, but we have selected five types of movement in a highway environment with relatively few objects, such as signal light and pedestrians (Table 3). In addition, AVs are carried out on the assumption that they do not cause injury to nearby vehicles (Sect. 3.2), and since the mixed situations between AVs and HVs are expected even after commercialization, the neighboring vehicles are set as HVs.

Table 3 Assumption for extracting scenarios

Environment	Driving behavior
2 lanes' expressway which has mixed between AVs and HVs	Proceeding straight (PS)
	Slowing (S)
	Passing other vehicle (POV)
	Changing lanes (CL)
	Merging (M)

Considering the MPC of AVs and the MPC of HVs, the combination of 25 scenarios (5 * 5) is derived, but considering the location and lane of AVs and HVs, a myriad of scenarios (25 * the location * the lanes) are derived. Therefore, we simulate the conditions separated by locational positions (ahead/rear) of AVs in two lanes environment. Over a hundred scenarios [25 * 2(ahead/rear) * 2(two lanes)] are generated which are presented in the order 'AVs (Ego)-HVs (Neighboring)' with each given code (movement preceding collisions are divided into PS, S, POV, CL, and M; Table 3).

4.2 AVs Ahead Situation

If AVs are forward vehicles, the surrounding vehicle, which is likely to crash, runs behind the AVs. Cases, where an accident may occur accordingly, shall be as shown in Table 4.

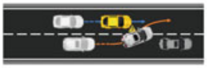
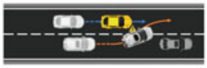
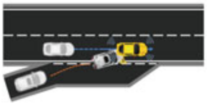
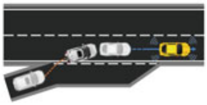
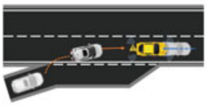
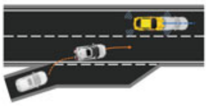
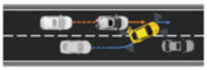
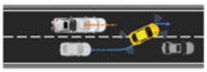
As a result of the scenario, it was found that the majority of scenarios where rear-end collisions occur depend on the behavior of the following neighborhood vehicles because the AVs are the leading vehicle. These scenarios have been found to occur a lot in passing other vehicles, changing lanes, merging situations where there is a lot of interaction between vehicles.

4.3 AVs Rear Situation

If AVs are rear vehicles, the surrounding vehicle, which is likely to crash, runs in front of the AVs. Cases, where an accident may occur accordingly, are shown in Table 5.

As a result of the scenario, since the Ego vehicle is a rear vehicle, it is expected that AVs maintain an environment where AVs do not crash according to the behavior of the preceding neighboring vehicle. Most traffic accidents occur in situations where there are some interactions such as passing other vehicles, changing lanes, and merging, and it is a frontal collision.

Table 4 Scenarios in AVs on ahead situation: it means that yellow car is the AVs, gray car is HVs, and warning sign is crash probability

Environment	Description	
	In the another lane	In the same lane
PS-PS		
PS-S		
PS-POV		
PS-CL		
PS-M		
S-PS		
S-S		
S-POV		
S-CL		
S-M		
POV-PS		
POV-S		
POV-POV		

(continued)

Table 4 (continued)

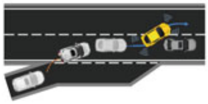
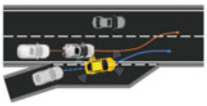
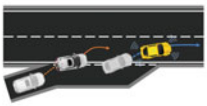
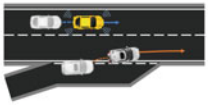
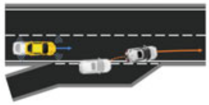
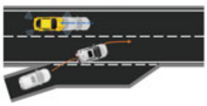
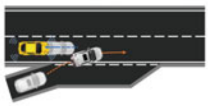
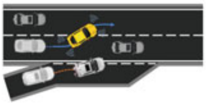
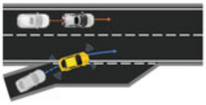
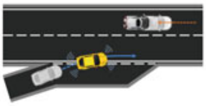
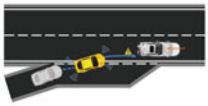
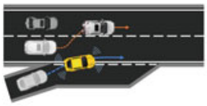
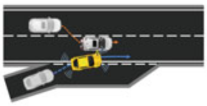
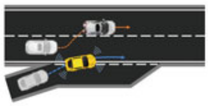
POV-CL	
POV-M	
CL-PS	
CL-S	
CL-POV	
CL-CL	
CL-M	
M-PS	
M-S	
M-POV	
M-CL	
M-M	

Table 5 Scenarios in AVs in the rear situation: it means that yellow car is the AVs, gray car is HVs, and warning sign is crash probability

Environment	Description	
	In the another lane	In the same lane
PS-PS		
PS-S		
PS-POV		
PS-CL		
PS-M		
S-PS		
S-S		
S-POV		
S-CL		
S-M		
POV-PS		
POV-S		
POV-POV		

(continued)

Table 5 (continued)

POV-CL		
POV-M		
CL-PS		
CL-S		
CL-POV		
CL-CL		
CL-M		
M-PS		
M-S		
M-POV		
M-CL		
M-M		

4.4 Comparison with HVs Traffic Accidents

NHTSA [34] states that 'The safety benefits of automated vehicles are par-amount. Automated vehicles' potential to save lives and reduce injuries is rooted in one critical and tragic fact: 94% of serious crashes are due to human error.' Prior to the introduction of AVs, physical accident factors between HVs are speed, type of vehicles, time, violation of the law, etc. [10, 35–39]. Even in mixed situations with HVs after AVs are introduced, the majority of accident factors will be caused by HVs, so the factors that determine the accident are probably not expected to change much. However, the introduction of AVs is expected to lead to making the new accident factors. For example, a human driver was aware of AVs, but it is expected that an accident could occur because the human driver himself judged it to be a risk factor for an accident [40]. Furthermore, it is expected that the increasing market penetration rate (MPR) of AVs could lead to unavoidable accident situations between AVs. Although there is a limit to analyzing current data, it needs to be carried out in a way that increases safety by identifying and preventing it in advance through various methods.

4.5 Summary

As a result of the overall scenario elicitation, accidents will vary depending on whether AVs are at the forefront or at the rear of the surrounding vehicles. In addition, different collisions may occur in the same situation depending on the lanes of the vehicle being operated. In particular, various accidents are expected to occur in the vicinity of passing lanes and ramp sections where there are many interactions. According to the regulation that the vehicles can only change to the left lane in the POV environment, many side collisions are expected, and it is leading to serious accidents depending on the high-speed operation. In addition, for CL environments, various accidents expected than POV environments due to the free lane change situation. In the M environment, it is the only entering traffic environment on the highway, and more accidents are expected due to the interaction between the entry vehicle and the mainline vehicle. Although these would be similar to the accident situation and type of accident between HVs currently operating, we only derived scenarios consisting of accidents caused by surrounding vehicles because AVs did not cause accidents.

5 Conclusion and Future Research

While various studies are being conducted on the safety of AVs, research has recently focused on the development of active safety systems. Among them, the field of AVs scenarios that directly affect vehicle safety and testing is actively being studied. However, while countless scenarios are expected to be derived from the process of scenario-making, the tendency to rely on expert knowledge and the lack of data utilization are emerging as problems. Accordingly, the study is being conducted to extract scenarios through minimizing dependencies and utilizing real data. As part of this flow, we have derived scenarios through machine learning techniques using about 300 data obtained from real traffic accidents of AVs.

As a result of eliciting determinants of AVs accidents, the MPC of HVs was identified as the most important factor, followed by the MPC of AVs. Based on this (most importantly), we have drawn accident situation scenarios that probably occur in various ways through the extracted vehicle-to-vehicle movement variables. The scenarios were shown up where the most interactions occur and resulted in a number of collisions at the overtaking, lane changes, and merging in the expressway. Such a collision is expected to be similar to that between HVs, but it is considered that there will be a difference in the degree of accident depending on the condition that AVs did not cause accidents. This work is expected to provide practical help in establishing scenarios by deriving basic scenarios through machine learning based on AVs real-world accident data and to mitigate expert knowledge dependencies.

Machine learning and AI require a larger amount of data, but in this study, there was a limitation of the lack of data. However, since this is expected to accumulate more autonomous vehicle accident data, this research could be expected to use as basic data in the future. Similarly, we disregarded many situations as scenarios derived from the assumptions of various situations. In addition, since AVs do not interact with only one vehicle, further studies are needed later considering this. It does not mean that AVs can only prevent accidents if they are rear than the surrounding vehicles, but imply that various/basic scenarios can be established by extracting accident determinants based on accident data. In the future, we will be presenting a study to establish scenarios in the multilane expressway and in urban environments where various objects interact with each other. Furthermore, beyond the presentation of simple scenarios, predictions of where accidents will occur from the Ego vehicle's perspective will directly help improve safety, so we will conduct these studies later.

Acknowledgements This work is supported by Ministry of Land, Infrastructure and Transport of Korea (grant 21AMDP-C161754-01) and 2020 Hongik University Research Fund.

References

1. https://time.com/3719270/you-asked-how-do-driverless-cars-work/. Access 03.04.2021
2. Navigant Leaderboard (2020) https://www.greencarcongress.com/2020/03/20200324-navigant.html
3. Abraham H, Lee C, Brady S, Fitzgerald C, Mehler B, Reimer B, Coughlin JF (2017) Autonomous vehicles and alternatives to driving: trust, preferences, and effects of age. In: Proceedings of the transportation research board 96th annual meeting

4. Zhang T, Tao D, Qu X, Zhang X, Lin R, Zhang W (2019) The roles of initial trust and perceived risk in public's acceptance of automated vehicles. Transp Res Part C Emerg Technol 98:207–220
5. Hartwich F, Witzlack C, Beggiato M, Krems JF (2019) The first impression counts—a combined driving simulator and test track study on the development of trust and acceptance of highly automated driving. Transp Res Part F Traff Psychol Behav 65:522–535
6. Goodall NJ (2014) Ethical decision making during automated vehicle crashes. Transp Res Rec 2424(1):58–65
7. Riedmaier S, Ponn T, Ludwig D, Schick B, Diermeyer F (2020) Survey on scenario-based safety assessment of automated vehicles. IEEE Access 8:87456–87477
8. Kim K, Kim B, Lee K, Ko B, Yi K (2017) Design of integrated risk management-based dynamic driving control of automated vehicles. IEEE Intell Transp Syst Mag 9(1):57–73
9. Kang M, Song J, Hwang K (2020) For preventative automated driving system (PADS): traffic accident context analysis based on deep neural networks. Electronics 9(11):1829
10. Lee H, Kang M, Song J, Hwang K (2020) The detection of black ice accidents for preventative automated vehicles using convolutional neural networks. Electronics 9(12):2178
11. Wen M, Park J, Cho K (2020) A scenario generation pipeline for autonomous vehicle simulators. HCIS 10:1–15
12. Ulbrich S, Menzel T, Reschka A, Schuldt F, Maurer M (2015) Defining and substantiating the terms scene, situation, and scenario for automated driving. In: 2015 IEEE 18th international conference on intelligent transportation systems. IEEE, pp 982–988
13. Webb N, Smith D, Ludwick C, Victor T, Hommes Q, Favaro F, Ivanov G, Daniel T (2020) Waymo. https://time.com/3719270/you-asked-how-do-driverless-cars-work/. Access: 03.04.2021's safety methodologies and safety readiness determinations. arXiv:2011.00054
14. Stellet JE, Zofka MR, Schumacher J, Schamm T, Niewels F, Zöllner JM (2015) Testing of advanced driver assistance towards automated driving: a survey and taxonomy on existing approaches and open questions. In: 2015 IEEE 18th international conference on intelligent transportation systems. IEEE, pp 1455–1462
15. De Gelder E, Paardekooper JP, Saberi AK, Elrofai H, Ploeg J, Friedmann L, De Schutter B (2020) Ontology for scenarios for the assessment of automated vehicles. arXiv:2001.11507
16. Schwall M, Daniel T, Victor T, Favaro F, Hohnhold H (2020) Waymo public road safety performance data. arXiv:2011.00038
17. Elrofai H, Paardekooper JP, de Gelder E, Kalisvaart S, den Camp OO (2018) Scenario-based safety validation of connected and automated driving. Netherlands Organization for Applied Scientific Research, TNO, Technical Report
18. Pütz A, Zlocki A, Bock J, Eckstein L (2017). System validation of highly automated vehicles with a database of relevant traffic scenarios. Situations 1:E5
19. Erdogan A, Ugranli B, Adali E, Sentas A, Mungan E, Kaplan E, Leitner A (2019) Real-world maneuver extraction for autonomous vehicle validation: a comparative study. In: 2019 IEEE intelligent vehicles symposium (IV). IEEE, pp 267–272
20. Webb N, Smith D, Ludwick C, Victor T, Hommes Q, Favaro F, Ivanov G, Daniel T (2020) Waymo's Safety Methodologies and Safety Readiness Determinations. arXiv:2011.00054
21. https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports/
22. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
23. Nadimi N, Ragland DR, Mohammadian Amiri A (2020) An evaluation of time-to-collision as a surrogate safety measure and a proposal of a new method for its application in safety analysis. Transp Lett 12(7):491–500
24. Li Y, Wu D, Lee J, Yang M, Shi Y (2020) Analysis of the transition condition of rear-end collisions using time-to-collision index and vehicle trajectory data. Accident Anal Prevent 144:105676
25. https://time.com/5205767/uber-autonomous-car-crash-arizona/. Access 10.03.2021
26. Lund UJ (2005) The effect of seating location on the injury of properly restrained children in child safety seats. Accid Anal Prev 37(3):435–439