Lecture Notes in Networks and Systems 945

Tomonobu Senjyu
Chakchai So–In
Amit Joshi *Editors*

Smart Trends in Computing and Communications

Proceedings of SmartCom 2024, Volume 1

Springer

Lecture Notes in Networks and Systems

Volume 945

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Türkiye

Derong Liu, Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

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

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Tomonobu Senjyu · Chakchai So–In · Amit Joshi
Editors

Smart Trends in Computing and Communications

Proceedings of SmartCom 2024, Volume 1

Springer

Editors
Tomonobu Senjyu
Faculty of Engineering
University of the Ryukyus
Nishihara, Okinawa, Japan

Chakchai So–In
College of Computing
Khon Kaen University
Khon Kaen, Thailand

Amit Joshi
Global Knowledge Research Foundation
Ahmedabad, Gujarat, India

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-97-1319-6 ISBN 978-981-97-1320-2 (eBook)
https://doi.org/10.1007/978-981-97-1320-2

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Paper in this product is recyclable.

Preface

The Eighth Edition of the SmartCom 2024—Smart Trends in Computing and Communications will be held during January 12 and 13, 2024, Physically at Hotel: Crowne Plaza Pune City Centre, Pune, India, and Digitally on Zoom which is organized by Global Knowledge Research Foundation and managed by G R Scholastic LLP. The associated partners were Springer Nature and Springer. National Chamber Partner Knowledge Chamber of Commerce and Industry. The conference will provide a useful and wide platform both for display of the latest research and for exchange of research results and thoughts. The participants of the conference will be from almost every part of the world, with background of either academia or industry, allowing a real multinational multicultural exchange of experiences and ideas.

A great pool of more than 1200 papers were received for this conference from across 18 countries among which around 202 papers were accepted with this springer series and were presented through digital platform during the two days. Due to overwhelming response, we had to drop many papers in hierarchy of the quality. A total of 33 technical sessions will be organized in parallel in two days along with few keynotes and panel discussions. The conference will be involved in deep discussion and issues which will be intended to solve at global levels. New technologies will be proposed, experiences will be shared, and future solutions for enhancement in systems and security will also be discussed. The final papers will be published in proceedings by Springer LNNS Series.

Over the years, this conference has been organized and conceptualized with collective efforts of a large number of individuals. I would like to thank each of the committee members and the reviewers for their excellent work in reviewing the papers. Grateful acknowledgements are extended to the team of Global Knowledge Research Foundation for their valuable efforts and support.

Nishihara, Japan — Tomonobu Senjyu
Khon Kaen, Thailand — Chakchai So–In
Ahmedabad, India — Amit Joshi

Contents

Editors and Contributors

About the Editors

Prof. Tomonobu Senjyu received his B.S. and M.S. degrees in Electrical Engineering from the University of the Ryukyus, Nishihara, Japan, in 1986 and 1988, respectively, and his Ph.D. in Electrical Engineering from Nagoya University, Japan, in 1994. He is currently Full Professor at the Department of Electrical and Electronics Engineering, University of the Ryukyus. His research interests include renewable energy, power system optimization and operation, power electronics, and advanced control of electrical machines.

Prof. Chakchai So–In has been with the Department of Computer Science at Khon Kaen University since 2010. He received his B.E. and M.E. degrees from Kasetsart University (KU), Bangkok, Thailand, in 1999 and 2001, respectively. He also completed M.S. and Ph.D. degrees in Computer Engineering at the Department of Computer Science and Engineering, Washington University, in St. Louis (WUSTL), MO, USA, in 2006 and 2010. He has authored more than 100 publications in prominent journals and proceedings, together with 10 books. His research interests include mobile computing/sensor networks, Internet of Things, computer/wireless/distributed networks, cyber security, intelligent systems, and the future Internet. He is Senior Member of the IEEE and ACM.

Dr. Amit Joshi is currently Director of Global Knowledge Research Foundation, also Entrepreneur Researcher who has completed his Masters' and research in the areas of cloud computing and cryptography in medical imaging. He has an experience of around 10 years in academic and industry in prestigious organizations. He is Active Member of ACM, IEEE, CSI, AMIE, IACSIT-Singapore, IDES, ACEEE, NPA, and many other professional societies. Currently, he is International Chair of InterYIT at International Federation of Information Processing (IFIP, Austria), He has presented and published more than 50 papers in national and international journals/conferences of IEEE and ACM. He has also edited more than 40 books which are published

by Springer, ACM and other reputed publishers. He has also organized more than 50 national and international conferences and programs in association with ACM, Springer, IEEE to name a few across different countries including India, UK, Europe, USA, Canada, Thailand, Egypt, and many more.

Contributors

K. R. Akilesh Department of CSE, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India

R. Angeline Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

Mrunal Annadate Dr. Vishwanath Karad MIT World Peace University, Kothrud, Pune, Bharat, India

Prutha Annadate Dr. Vishwanath Karad MIT World Peace University, Kothrud, Pune, Bharat, India

Linu Tess Antony Amal Jyothi College of Engineering, Koovappally, APJAKTU, Trivandrum, Kerala, India

J. Sangeetha Archana Department of CSE, Centre of Excellence in Assistive Technology, Rajalakshmi Engineering College, Chennai, India

Rupali Arora University School of Business, Chandigarh University, Mohali, Punjab, India

R. C. Arya Department of CSE, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India

Poonam Bansal Department of Artificial Intelligence and Data Sciences, Indira Gandhi Delhi Technical University for Women (IGDTUW), Delhi, India

Mangesh Bedekar Dr. Vishwanath Karad MIT World Peace University, Kothrud, Pune, Bharat, India

M. A. Bharathi BMS Institute of Technology and Management, Bengaluru, India

R. Bharathi PES University, Bangalore, India

S. Vijayakumar Bharathi Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, India

Kriti Bhushan National Institute of Technology, Kurukshetra, India

Prasun Chakrabarti Department of Computer Science and Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India

Nilotpal Chakraborty Indian Institute of Information Technology Guwahati, Guwahati, India

Somyajit Chakraborty Indian Institute of Information Technology Guwahati, Guwahati, India

Satakshi Chatterjee Amity School of Business, Amity University Jharkhand, Nivaranpur, Ranchi, Jharkhand, India

Manasa S. Desai Siddaganga Institute of Technology, Tumakuru, India

Amita Dev Indira Gandhi Delhi Technical University for Women, Delhi, India

Rugved Dharmadhikari Pune Institute of Computer Technology, Pune, India

Mukta Dhopeshwarkar Department of CS and IT, Dr. Babasaheb Ambedkar, Marathwada University, Chhatrapati Sambhajinagar, India

Sanika Gaikwad Pune Institute of Computer Technology, Pune, India

Avijit Gayen Indian Institute of Information Technology Guwahati, Guwahati, India

Rita Ghial University Institute of Legal Studies, Panjab University, Chandigarh, India

Arunangshu Giri School of Management and Social Sciences, Haldia Institute of Technology, ICARE Complex, Hatiberia, Haldia, West Bengal, India

Neetika Goel Department of Artificial Intelligence and Data Sciences, IGDTUW, Delhi, India

Sudha Govindan Part Time Research Scholar, Department of Computer Science, Madurai Kamaraj University, Madurai, Tamil Nadu, India

Deepti Gupta Department of Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, India

Tanish Gupta Pune Institute of Computer Technology, Pune, India

Sachal Hablani Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Sidharth Raja Halder Amity School of Business, Amity University Jharkhand, Nivaranpur, Ranchi, Jharkhand, India

Kaustubh Hatkar SCTR's Pune Institute of Computer Technology, Pune, India

Bhawna Jain Department of Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, India

R. C. Jaiswal Pune Institute of Computer Technology, Pune, India;
SCTR's Pune Institute of Computer Technology, Pune, India

Angshuman Jana Indian Institute of Information Technology Guwahati, Guwahati, India

Shubham Jha Siddaganga Institute of Technology, Tumakuru, India

Manjusha Joshi Amity University, Panvel, Maharashtra, India

Pranjali Joshi Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

A. Karnan Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Annamalainagar, Tamil Nadu, India

Abhishek Katke Pune Institute of Computer Technology, Pune, India

Rajinder Kaur University Institute of Legal Studies, Panjab University, Chandigarh, India

Aritra Khasnobis National Institute of Technology, Kurukshetra, India

Vedant Khergade Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Monali Korde Department of Artificial Intelligence and Machine Learning, A. P. Shah Institute of Technology, Thane, India

Manjunath Sargur Krishnamurthy JP Morgan & Chase Co., Bengaluru, India

Khyati Kulkarni Pune Institute of Computer Technology, Pune, India

Varad Kulkarni Pune Institute of Computer Technology, Pune, India

Peeyush Kulshrestha Department of Computer Science Engineering, Poornima Institute of Engineering and Technology, Jaipur, India

S. N. Kumar Amal Jyothi College of Engineering, Koovappally, APJAKTU, Trivandrum, Kerala, India

Tarun Kumar CPDM, Indian Institute of Science, Bengaluru, India

Kavita Kumari University School of Business, Chandigarh University, Mohali, Punjab, India

Ritika Kumari Department of Artificial Intelligence and Data Sciences, Indira Gandhi Delhi Technical University for Women (IGDTUW), Delhi, India; USICT, Guru Gobind Singh Indraprastha University, New Delhi, India

Gaurav Kumawat Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, Rajasthan, India

Pranjal Lamsal Siddaganga Institute of Technology, Tumakuru, India

Vivian Brian Lobo TSEC, Bandra, Mumbai, India

Sanket Lokhande SCTR's Pune Institute of Computer Technology, Pune, India

C. Maithri Kalpataru Institute of Technology, Tiptur, India

Fredy Malcom Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

Sudhir Manivannan PES University, Bangalore, India

Gabriela Mogos School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China

Nipun Morayya Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Zain Muhammed PES University, Bangalore, India

Aarya Mujumdar Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Mousami V. Munot SCTR's Pune Institute of Computer Technology, Pune, India

M. B. Nirmala Siddaganga Institute of Technology, Tumakuru, India

G. Nityasree Department of CSE, Centre of Excellence in Assistive Technology, Rajalakshmi Engineering College, Chennai, India

Harsh Patani Pune Institute of Computer Technology, Pune, India

Kanchan Pranay Patil Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, India

Soham Lalit Patil Department of Computer Engineering, Modern Education Society's College of Engineering, Pune, India

Maitreyi Phadke Department of Artificial Intelligence and Machine Learning, A. P. Shah Institute of Technology, Thane, India

H. Pooja PES University, Bangalore, India

S. Poonkuzhali Department of CSE, Centre of Excellence in Assistive Technology, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India

Yuvan Pradhan PES University, Bangalore, India

Pijush Kanti Dutta Pramanik School of Computer Application and Technology, Galgotias University, Greater Noida, India

Dhanya Pramod Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, India

K. K. Prashanthini Department of CSE, Centre of Excellence in Assistive Technology, Rajalakshmi Engineering College, Chennai, India

Aprajita Priya Siddaganga Institute of Technology, Tumakuru, India

Diksha Pundir University Institute of Legal Studies, Panjab University, Chandigarh, India

K. Rachitha Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

R. Ragupathy Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Annamalainagar, Tamil Nadu, India

Manish Prasad Rajak Amity School of Business, Amity University Jharkhand, Nivaranpur, Ranchi, Jharkhand, India

I. S. Rajesh BMS Institute of Technology and Management, Bengaluru, India

Anagha Rajput Pune Institute of Computer Technology, Pune, India

S. B. Rammya PES University, Bengaluru, Karnataka, India

Madhuri Rao TSEC, Bandra, Mumbai, India

Bhavikkumar Ronvelia Electronics and Communication Department Institute of Technology, Nirma University, Ahmedabad, Gujarat, India

Priyanka Roy Department of Electronics and Communication Engineering, Tripura Institute of Technology, Narsingarh, Tripura, India

Kasturi Sahay Amity School of Business, Amity University Jharkhand, Nivaranpur, Ranchi, Jharkhand, India

Kanhu Charan Sahoo SCTR's Pune Institute of Computer Technology, Pune, India

Suguna Sangaiah Associate Professor, PG and Research Department of Computer Science, Sri Meenakshi Government Arts College for Women(A), Madurai, Tamil Nadu, India

M. G. Sathyaa Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

Vaidehi Satushe Department of Electronics and Telecommunication Engineering, COEP Technological University (COEP Tech), Pune, Maharashtra, India

Rituparna Sen Department of Electronics and Communication Engineering, Tripura Institute of Technology, Narsingarh, Tripura, India

Dhaval Shah Electronics and Communication Department Institute of Technology, Nirma University, Ahmedabad, Gujarat, India

D. Sorna Shanthi Department of CSE, Centre of Excellence in Assistive Technology, Rajalakshmi Engineering College, Chennai, India

Aniva Sharma Department of Electrical Engineering, Poornima Institute of Engineering and Technology, Jaipur, India

Arun Sharma Department of Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, India

Om Shinde Pune Institute of Computer Technology, Pune, India

Ulka Shirole A. C. Patil College of Engineering, Navi Mumbai, Maharashtra, India

S. Shivkumar Department of Computer Science Engineering and Amrita School of Computing, Amrita Vishwa Vidyapeetham, Bengaluru, Karnataka, India

Adityapratap Singh Pune Institute of Computer Technology, Pune, India

Mayank Raj Singh Siddaganga Institute of Technology, Tumakuru, India

Pravinkumar Sonsare Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

S. Soorya Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

J. Sowmiya Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

M. Supriya Department of Computer Science Engineering and Amrita School of Computing, Amrita Vishwa Vidyapeetham, Bengaluru, Karnataka, India

Sarika Swami Department of CS and IT, Dr. Babasaheb Ambedkar, Marathwada University, Chhatrapati Sambhajinagar, India

Tamanna Indira Gandhi Delhi Technical University for Women, Delhi, India

Vansh Teppalwar SCTR's Pune Institute of Computer Technology, Pune, India

U. K. Tharun Suriya Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

Puja Tiwari Department of Management Studies, Haldia Institute of Management, ICARE Complex, Hatiberia, Haldia, West Bengal, India

Bijoy Kumar Upadhyaya Department of Electronics and Communication Engineering, Tripura Institute of Technology, Narsingarh, Tripura, India

Priya Vijay Department of CSE, Centre of Excellence in Assistive Technology, Rajalakshmi Engineering College, Chennai, India

Santosh Kumar Vishwakarma Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, Rajasthan, India

Vibha Vyas Department of Electronics and Telecommunication Engineering, COEP Technological University (COEP Tech), Pune, Maharashtra, India

Cancer Classification Using Gene Expression Data

Pravinkumar Sonsare, Aarya Mujumdar, Pranjali Joshi, Nipun Morayya, Sachal Hablani, and Vedant Khergade

Abstract Cancer poses a significant global health challenge, necessitating personalized therapeutic approaches. Genomic prediction, utilizing large-scale genomic data, has emerged as a powerful tool in cancer research. The proposed effort aims to create reliable machine learning models for cancer classification with the use of the UCI cancer dataset. The models will identify intricate patterns and correlations, enabling accurate classification of cancer subtypes and providing personalized treatment recommendations by integrating genomic and clinical data. The dataset is analyzed using machine learning techniques like k-nearest neighbors, support vector machines, random forest, and logistic regression. We found that k-neighbors classifier has an accuracy of 94.9, support vector machine has 95.6, random forest has 95.2, and logistic regression has the highest accuracy of 95.8.

Keywords Large-scale genomic data · Cancer research · Machine learning models · Cancer subtypes · KNN · SVM · RF · Logistic regression · Gene expression

P. Sonsare (✉) · A. Mujumdar · P. Joshi · N. Morayya · S. Hablani · V. Khergade
Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India
e-mail: sonsarep@rknec.edu

A. Mujumdar
e-mail: mujumdaraa@rknec.edu

P. Joshi
e-mail: joshipu_1@rknec.edu

N. Morayya
e-mail: morayyana@rknec.edu

S. Hablani
e-mail: hablanisr_1@rknec.edu

V. Khergade
e-mail: khergadevs@rknec.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

T. Senjyu et al. (eds.), *Smart Trends in Computing and Communications*, Lecture Notes in Networks and Systems 945, https://doi.org/10.1007/978-981-97-1320-2_1

1 Introduction

Cancer remains a global health challenge, imposing a substantial burden on individuals, families, and healthcare systems. Despite significant advancements in understanding cancer's molecular basis, the complex interplay of hereditary and environmental factors leads to diverse treatment outcomes among patients. Genomic prediction has emerged as a crucial tool in cancer research, utilizing genomic data to identify genetic variations influencing disease susceptibility, prognosis, and treatment response.

Genomic prediction incorporates high-throughput techniques like RNA sequencing to measure gene expression levels, providing functional insights into how genetic variations impact cellular processes. Integrating gene expression data into prediction models enhances their accuracy, helping identify gene sets associated with cancer subtypes, progression stages, and treatment outcomes. Machine learning algorithms, including support vector machines, random forests, and deep learning models, leverage gene expression profiles to build accurate prediction models.

The challenge in genomic prediction involves developing machine learning models that predict an individual's cancer risk using their genomic data. In this work, our goal is to create robust models leveraging large genomic datasets, identifying intricate patterns and correlations challenging for human prediction. These models will be trained on comprehensive genomic data, encompassing genetic variations and clinical factors, aiming to classify cancer types, identify risk factors, and provide personalized treatment suggestions.

Machine learning allows us to establish connections between diverse variables and diseases. This work explores data processing, representation, preparation, and feature selection methods. Using the UCI cancer dataset with rich genomic and clinical data from 800 patients, our objective is to construct a machine learning model for cancer classification.

To achieve this, we adopt a machine learning approach, leveraging algorithms like k-nearest neighbors, support vector machines, random forest, and logistic regression. These algorithms, proven effective in genomic prediction studies, will be trained on a subset of the UCI data, incorporating both genomic features and clinical variables. Analyzing this combined information, the model aims to accurately classify cancer samples, predict cancer subtypes, and guide treatment decisions.

2 Literature Review

Numerous studies have been done on genomic prediction in cancer or risk of cancer. They have applied different techniques for detection and achieved different accuracies for different algorithms. Several methods that combined ML techniques with effective feature engineering and transfer learning methods have produced results with good test accuracies. Table 1 depicts a brief description of various models.

Table 1 Consolidated review of different literature papers

References	Dataset	Algorithm	Dataset type	Performance
Segal et al. [4]	Cancer	SVM	Gene expression data	Accuracy: 98.5%
Hijazi et al. [5]	Mixed-lineage leukemia (MLL)	SVM linear	Gene expression data	Accuracy: 99.89%
Ram et al. [6]	Colon cancer	RF	Microarray data	Accuracy: 87.39%
Zhang et al. [7]	Breast cancer	SVM-RFE-PSO	Gene expression data	Accuracy: 81.54%
Yuan et al. [1]	Tumor-educated platelets	Evolutionary programming-trained SVM	Gene expression data	Accuracy: 95.93%
Yuan et al. [2]	Lung adenocarcinoma (AC) and lung squamous cell cancer (SCC)	RF	Gene expression data	Accuracy: 94.9%
		RF	Gene expression data	Accuracy: 93.3%
		SVM	Gene expression data	Accuracy: 94.7%
Abdulqader et al. [3]	Lymphoma	kNN	Microarray data	Accuracy: 94.7%
	Lymphoma	NB	Microarray data	Accuracy: 74.83%

The table shows a comprehensive analysis of cancer classification studies, showcasing diverse datasets, algorithms, dataset types, and their respective performance metrics. Notably, Hijazi et al. [5] achieved an exceptional 99.89% accuracy in identifying mixed-lineage leukemia (MLL) using SVM Linear. Yuan et al. [1, 2] demonstrated promising results, with accuracies of 95.93% for tumor-educated platelets and 94.9% for lung adenocarcinoma using different algorithms. Conversely, Abdulqader et al. [3] reported varying outcomes, achieving 94.7% accuracy for lymphoma with KNN but a lower 74.83% using NB. This analysis highlights the wide-ranging methods and varying success rates in applying machine learning for cancer diagnosis and classification, underscoring the complexity and challenges of the field.

3 Datasets

The dataset is publicly available for free on the UCI Machine Learning repository and can be downloaded in CSV format. This dataset contains data on gene expression from RNA sequencing (RNA-Seq) research and is frequently utilized in cancer research. Samples, also referred to as instances, are organized in rows. For each sample, the variables or attributes consist of RNA-Seq gene expression levels, which are measured using the Illumina HiSeq platform. Gene expression levels in this dataset refer to the amount of RNA produced by a particular gene in a sample.

RNA-Seq is a technique used to measure the expression levels of genes in a sample. It provides valuable insights into the activity of genes and their potential role in various biological processes, including cancer. The dataset consists of a collection of samples from different cancer types, including breast cancer, kidney renal cancer, prostate adenocarcinoma, and others. Each sample represents a patient and contains information about the expression levels of thousands of genes.

The dataset consists of a total of 801 samples or instances. Each sample is described by 20,531 attributes, which correspond to the expression levels of different genes. These attributes are continuous numerical values. Each sample is labeled with a specific cancer type, such as (“breast invasive carcinoma” (BRCA), “colon adenocarcinoma” (COAD), (“lung adenocarcinoma” (LUAD), “prostrate adenocarcinoma” (PRAD), and “kidney renal cancer” (KIRC). These class labels represent the target variable that is used for classification tasks.

4 Proposed Work

4.1 The Various Phases Are

Data Preprocessing

Data preprocessing in machine learning is a pivotal stage that elevates the data’s quality, enabling the extraction of valuable insights. Real-world data often arrives incomplete, inconsistent, inaccurate, and with missing attribute values. Preprocessing serves as a data mining technique, converting raw data into a comprehensible format, thus preparing it for the construction and training of machine learning models.

Initially, Python libraries dedicated to preprocessing tasks are imported. The three foundational Python libraries employed for data preprocessing in machine learning are NumPy, Pandas, and Matplotlib. NumPy facilitates the conversion of data frames into NumPy arrays, while Pandas supports data manipulation, analysis, dataset management, and importation. Matplotlib, a 2D plotting library, aids in visualizing data through charts.

Subsequently, the dataset is imported using Pandas. Every machine learning dataset must be split into a training set and a test set, two separate groups. The model is trained with the help of the training set, leveraging known outputs. On the other hand, the test set is utilized for outcome prediction.

Classification Modeling
After the preprocessing of data, we give this modified dataset to the four machine learning algorithms and make further predictions. We will use Python and Jupyter Notebook. We will visualize the dataset and find the accuracy in the prediction of the system. A general workflow of the prediction will be created, and a correlation between features will be found. Predictions will be made using machine learning algorithms.

4.2 The Four ML Algorithms Are

K-Nearest Neighbors
The algorithm utilizes the Euclidean distance function, denoted as dis(a, b), to extract information from the samples. This is accomplished by employing the majority vote of the k-nearest neighbors. Equation (1) shows how distance is calculated,

$$\text{Dis}\,(a,b) = \sqrt{\left((a_1 - b_1)^2 + \cdots + (a_n - b_n)^2\right)}$$

$$= \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}. \tag{1}$$

Support Vector Machines
SVM locates the hyperplane by using the samples and the below equation,

$$f(X) = w^T x + b, \tag{2}$$

where

w is a vector that is normal to hyperplane

b is an intercept.

This hyperplane is also known as the optimal solution. The distance of a hyperplane equation from a given point vector x_0 can be written as:

$$d_h(x_0) = \frac{\left|w^T x_0 + b\right|}{||w||^2}. \tag{3}$$

Here $||w||^2$ is the Euclidean norm and is calculated as,

$$||w||^2 = \sqrt{w_1^2 + w_2^2 + w_3^2 + \ldots . w_n^2}. \tag{4}$$

Random Forest
This represents a bagging methodology that involves classifying samples through the construction of multiple decision trees, aimed at achieving the optimal outcome. Given a dataset D, where $P = \{p_1, p_2, p_3, \ldots, p_n\}$ are features and corresponding target values are $Q = \{q_1, q_2, q_3, \ldots, q_n\}$, the bagging procedure is carried out from $i = 1$ to N. To predict outcomes for new, unseen samples, the approach entails averaging the predictions across the p values. In other words, the predictions are aggregated to determine the final result.

$$y = \left(\frac{1}{N} \sum_{i=1}^{N} fb(p) \right). \tag{5}$$

By analyzing the individual trees' standard deviations, the error is calculated.

$$\text{S.D.} = \frac{\left(\sum_{i=1}^{nN} (fb(p) - f)^2 \right)}{2(N-1)}. \tag{6}$$

Logistic Regression
Logistic regression employs functions known as logit functions, which establish a connection between the dependent variable and independent variables by forecasting probabilities or likelihoods of events occurring.

The logistic functions, often referred to as sigmoid functions, transform these probabilities into binary values, facilitating their use for prediction purposes. Within logistic regression, the constant (b) shifts the curve horizontally, while the slope ($b1$) determines the degree of transformation's steepness.

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}. \tag{7}$$

5 Experimental Analysis

5.1 Import Libraries

When dealing with massive datasets, the efficient management, representation, and processing of data are essential tasks, and Python provides a set of standard libraries to address these challenges. Among these libraries, NumPy was used for array handling and numerical computations. For data manipulation and analysis, Pandas library proves indispensable, offering a versatile range of functions to efficiently handle structured data. Scikit-learn, commonly referred to as sklearn, was used for

implementing various machine learning algorithms and conducting in-depth statistical analysis. Another significant library used was TensorFlow. TensorFlow finds its application in normalizing data, a crucial preprocessing step that ensures data consistency and enhances model performance. Additionally, for effective data visualization, matplotlib was utilized, helping us create captivating figures and plots to gain insights from the data visually.

5.2 Import Dataset

The dataset is present in our system as "data.csv" and "labels.csv." We will make use of Pandas library's read_csv() function to get the dataset. The dataset gets loaded into the variable dataset. We can access and modify the dataset using this variable.

5.3 Data Processing and Splitting

After viewing the dataset, we used statistical methods of mean imputation to fill in missing values in gene expression data. Further, we concluded that we need to convert some categorical variables to numerical ones and normalize all values before we start training the ML models. We convert the categorical variables to numerical values. We normalized the dataset. Our dataset is not ready to be applied to machine learning algorithms. We separate the data into training and testing. We used four different algorithms: K-neighbors classifier, support vector machine, random forest classifier, and decision tree for experimentation.

5.4 Modeling

For the modeling, we will be using four different algorithms: K-neighbors classifier, support vector machine, random forest classifier, and decision tree.

K-Neighbors Classifier
The classification score varies based on the number of neighbors. Hence, we will plot a score graph for various values of K (neighbors) and check for the best score achieved. We have accuracy_scores[] to hold the accuracy scores of different neighbors and will plot the graph clearly to see which value of K gives the best score.

The maximum score that was achieved was 94.9 with $k = 9$ as shown in Fig. 1.

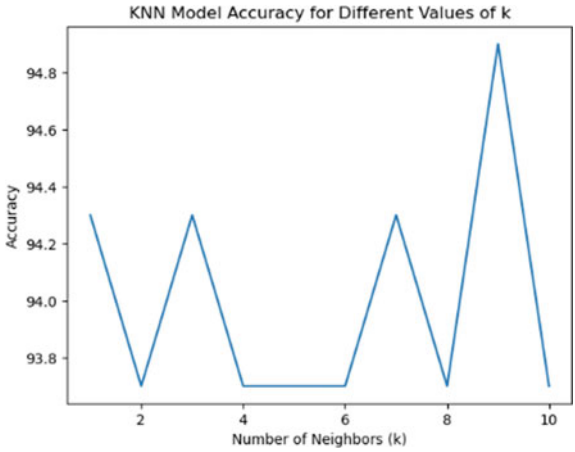

Fig. 1 K-neighbors classifier scores for different K values

Support Vector Machine

We first imported SVM from sklearn and then created a classifier which was later trained using training sets. Then predicted the results using this classifier. The accuracy achieved is 95.6%. The precision achieved is 95.3, and the recall score achieved is 95.1. The heatmap for various parameters is shown in Fig. 2.

Random Forest Classifier

We applied the random forest classifier for different values of estimators in the range [100, 120]. We appended the score for each respective number of estimators to find the maximum accuracy achieved by this classifier. The graph plotted for each number of estimators and their respective scores is shown in Fig. 3.

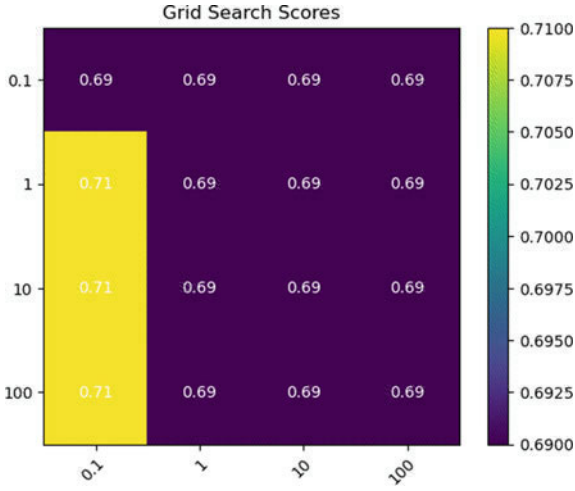

Fig. 2 Heatmap for different values of hyperparameters c, gamma, kernel

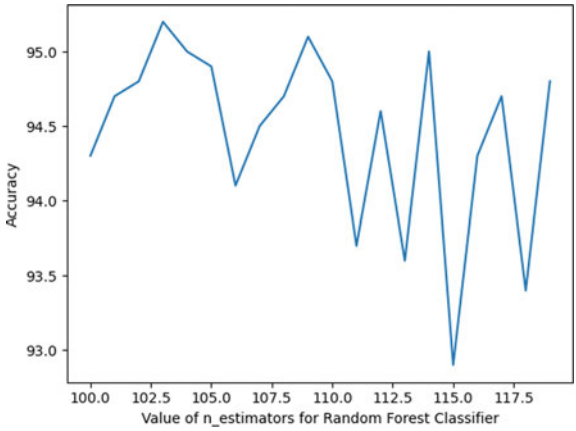

Fig. 3 Graph for random forest classifier for different values of estimators

The highest accuracy achieved for 103 estimators is 0.952. The precision achieved is 0.95.

Logistic Regression
We instantiated the logistic regression model with solver = 'lbfgs' and max_iter = 1000. After training the model, we predicted the results with an accuracy value of 95.8. The precision achieved is 95.4, and the recall score achieved is 95.

5.5 Model Selection and Application

We trained these models on NVIDIA IA DGX A100 with 640 GB total GPU memory. After exploring various models, it is time to select the best model that will be able to predict the results with the highest accuracy. Table 2 shows the accuracies achieved by the various algorithms we explored.

Further, let us look at their precision and recall values in Table 3.

We chose a model based on the computational complexity of the algorithm. Some models, like k-nearest neighbors (KNN), can be computationally expensive when dealing with large datasets. Models like logistic regression or random forest are more suitable when computational efficiency is a concern. As logistic regression has more accuracy than random forest, it is more suitable for this study.

Table 2 Accuracies for different algorithms

Algorithms	Accuracy (%)
K-nearest neighbors classifier	94.9
Support vector machine	95.6
Logistic regression	95.8
Random forest	95.2

Table 3 Precision and recall values for given algorithms

Algorithms	Precision (%)	Recall (%)
K-nearest neighbors classifier	94.2	94.3
Support vector machine	95.3	95.1
Logistic regression	95.4	95.5
Random forest	95	95

6 Result and Discussion

By training all four algorithms, i.e., k-nearest neighbors classifiers, support vector machine, logistic regression, random forest algorithms and computing the accuracy score, we got that k-neighbors classifiers have an accuracy of 94.9, support vector machine has 95.6, random forest has 95.2, and logistic regression has the highest accuracy of 95.8. We have computed accuracies by using method accuracy_score of library sklearn.metrics. Subset accuracy is calculated using the accuracy function, and for a sample to be accurate, the labels predicted for it must exactly match the equivalent set of labels in the ground truth (y_true). After selecting the model with the best accuracy, we trained it and tested on different data. We have built a system that takes in the gene expression values of patients' outputs if the patient has a certain type of cancer or not.

This study successfully predicted cancer, with 95.8% accuracy. The model used is logistic regression. This study can be extended to use some other model and diverse datasets if needed in future to train on models and predict the types of cancer. Furthermore, to address the challenge of interpretability, we plan to explore interpretable machine learning techniques and methodologies that allow us to extract meaningful insights from complex models.

Yuan et al. [2] employed SVM and random forest in incremental feature selection and tested on a different number of features. SVM achieved an accuracy of 94.9% and random forest achieved an accuracy of 94.7%.

7 Conclusion and Future Scope

In the past fifteen years, significant strides in cancer genomics have been marked by advancements in sequencing technology and robust computational algorithms. These developments played a crucial role in global projects focused on sequencing cancer genomes, making the process more efficient and cost-effective.

Gene expressions are acknowledged as pivotal in addressing core challenges in cancer diagnosis, treatment, and drug development. The introduction of microarray technology has facilitated the generation of large volumes of gene expression data. Looking ahead, the application of various deep learning methods is anticipated to enhance the dataset's accuracy.

References

1. Yuan L, Sun Y, Huang G (2020b) Using class-specific feature selection for cancer detection with gene expression profile data of platelets. Sensors 20:1528 [CrossRef]
2. Yuan F, Lu L, Zou Q (2020a) Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. Biochim Biophys Acta BBA Mol Basis Dis 1866:165822. [CrossRef] [PubMed]
3. Abdulqader DM, Abdulazeez AM, Zeebaree DQ (2020) Machine learning supervised algorithms of gene selection: a review. Mach Learn 62:233–244
4. Segal NH, Pavlidis P, Noble WS, Antonescu CR, Viale A, Wesley UV, Busam K, Gallardo H, DeSantis D, Brennan MF et al (2003) Classification of clear-cell sarcoma as a subtype of melanoma by genomic profiling. JCO 21:1775–1781. [CrossRef] [PubMed]
5. Hijazi H, Chan C (2013) A classification framework applied to cancer gene expression profiles. J Healthc Eng 4:255–284. [CrossRef]
6. Ram, M.; Najafi, A.; Shakeri, M.T. Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest. Iran J. Pathol. 2017, 12, 339–347. [CrossRef] [PubMed]
7. Zhang, Y.; Deng, Q.; Liang, W.; Zou, X. An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data. BioMed Res. Int. 2018, 2018, 7538204. [CrossRef] [PubMed]
8. Ferroni P, Zanzotto F, Riondino S, Scarpato N, Guadagni F, Roselli M (2019) Breast cancer prognosis using a machine learning approach. Cancers 11(3):328
9. Goli S, Mahjub H, Faradmal J, Mashayekhi H, Soltanian AR (2016) Survival prediction and feature selection in patients with breast cancer using support vector regression. Comput Math Methods Med, 2016
10. Akcay M, Etiz D, Celik O (2020) Prediction of survival and recurrence patterns by machine learning in gastric cancer cases undergoing radiation therapy and chemotherapy. Adv Radiat Oncol 5(6):1179–1187
11. Lu F, Chen Z, Yuan X, Li Q, Du Z, Luo L, Zhang F (2017) MMHG: multi-modal hypergraph learning for overall survival after D2 gastrectomy for gastric cancer. In: 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech). IEEE, pp 164–169
12. Momenzadeh N, Hafezalseheh H, Nayebpour MR, Fathian M, Noorossana R (2021) A hybrid machine learning approach for predicting survival of patients with prostate cancer: a SEER-based population study. Inf Med Unlocked 27:100763
13. Roy B, Stepišnik T, ALS TPROA, Vens C, Džeroski S, Clinical Trials Consortium (2022) Survival analysis with semi-supervised predictive clustering trees. Comput Biol Med 141:105001
14. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK (2019) Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med Inform Decis Mak 19:1–17
15. Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J (2019) Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clin Epidemiol Glob Health 7(3):293–299
16. Sarkar JP, Saha I, Sarkar A, Maulik U (2021) Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. Comput Biol Med 131:104244
17. Deng F, Zhou H, Lin Y, Heim JA, Shen L, Li Y, Zhang L (2021) Predict multicategory causes of death in lung cancer patients using clinicopathologic factors. Comput Biol Med 129:104161

A Three-Phase Heart Rate Variability Analysis for Assessing Cardiac Health Deterioration in Diabetics

Ulka Shirole and Manjusha Joshi

Abstract Diabetes is a prevalent global ailment that results in dysfunction in multiple organs throughout the body. Dysfunction in the autonomous nervous system (ANS) and cardiac health due to elevated glucose levels are studied from one time heart rate variability (HRV) analysis. Regular analysis is essential to monitor the rate of degradation caused by diabetes. This study introduces a three-stage HRV analysis to investigate the cardiac health decline resulting from diabetes. Diabetes data, collected during 15-minute recording sessions in both supine and sitting positions, undergoes analysis in three distinct phases. The time gap between the two ECG readings is six months. Patients with diabetes may or may not be taking medication. The outcomes of a three-point investigation will aid in evaluating changes, either deteriorative or ameliorative, in the cardiac health of these patients. The findings indicate that the orthostatic stress index (OSI) and the LF/HF ratio prove to be highly responsive and efficient indicators for assessing the decline in cardiac health. These parameter values can guide medical practitioners in prescribing appropriate medications.

Keywords Diabetes · Heart rate variability · Autonomous nervous system · Cardiac health

1 Introduction

Diabetes is a pervasive chronic condition that adversely affects multiple human organs, leading to complicated diseases that sometimes result in mortality [1]. Elevated blood glucose levels caused by diabetes lead to complications and the dys-

U. Shirole (✉)
A. C. Patil College of Engineering, Navi Mumbai 410210, Maharashtra, India
e-mail: umshirole@acpce.ac.in

M. Joshi
Amity University, Mumbai-Pune Express Way, Panvel 410206, Maharashtra, India
e-mail: mjmanjusha@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

T. Senjyu et al. (eds.), *Smart Trends in Computing and Communications*, Lecture Notes in Networks and Systems 945, https://doi.org/10.1007/978-981-97-1320-2_2

function of various body organs [2, 3]. If left untreated, diabetes can cause heart and blood vessel disease and sometime can cause heart attack [4]. According to the IDF Diabetes Atlas 9th edition 2019, in India, the prevalence of diabetes in adults is 11.3%. According to a study conducted by Dr. Mohan and reported in the Economic Times Health World, in 2018, India has a mortality rate that is three times greater as a result of diabetes [5]. The use of HRV analysis for assessing cardiac health has gained popularity and proven valuable due to its capacity to detect subtle changes in sympathetic and parasympathetic activity, offering insights into autonomic function [6].

In this experiment, a patient's cardiac health is assessed with the help of HRV data collected three times in both sitting and supine positions for 15 min, with a six-month time interval between two recordings. Diabetes patients may or may not be on medication. Changes in the HRV values from sitting to supine position will help to study the ability of the heart's response to increased stress. A three-point study will help to analyse whether there is a deterioration or improvement in the patient's cardiac health. Deterioration in cardiac health is associated with a decreased HRV parameter value [7, 8]. The decrease in HRV from the first to the second recording and from the second to the third can be used by the medical practitioner to advise the patient on the proper medication dosage.

2 Literature Survey

HRV has been assessed in many cardiac, neurological, and rheumatological illnesses in recent years and has come into the foreground as an essential marker of mortality. The heart rate modulation is controlled by various regulatory processes in the human body, such as the respiration system, endocrine system, the metabolism system, and the nervous system [9]. HRV measures alterations in an autonomic balance analogous to the onset of the particular disease and the variations observed during cardiac autonomic control. Researchers have also investigated the influence of HRV parameters under various physiological conditions [7, 10–12]. Numerous studies evidence that HRV is an effective and powerful means for the crucial measure of mortality [10–12]. HRV has been extensively researched in patients suffering from congestive coronary heart failure (CHF) [11, 12]. The influence of the autonomic nervous system modulation on the heart is assessed using HRV [7]. With the HRV technique, efferent cardiac sympathetic-parasympathetic modulation may be measured at the sinus node stage.

Selye [13] defined stress as the response to the stimulus that disturbs stability of the body and breaks mental and physical balance. Stress was also characterized as a detrimental condition in which the sympathetic nervous system (SNS) becomes excessively activated, resulting in acute or chronic physical, mental, and behavioural dysfunction [14].

Currently, there is no established benchmark for assessing stress.