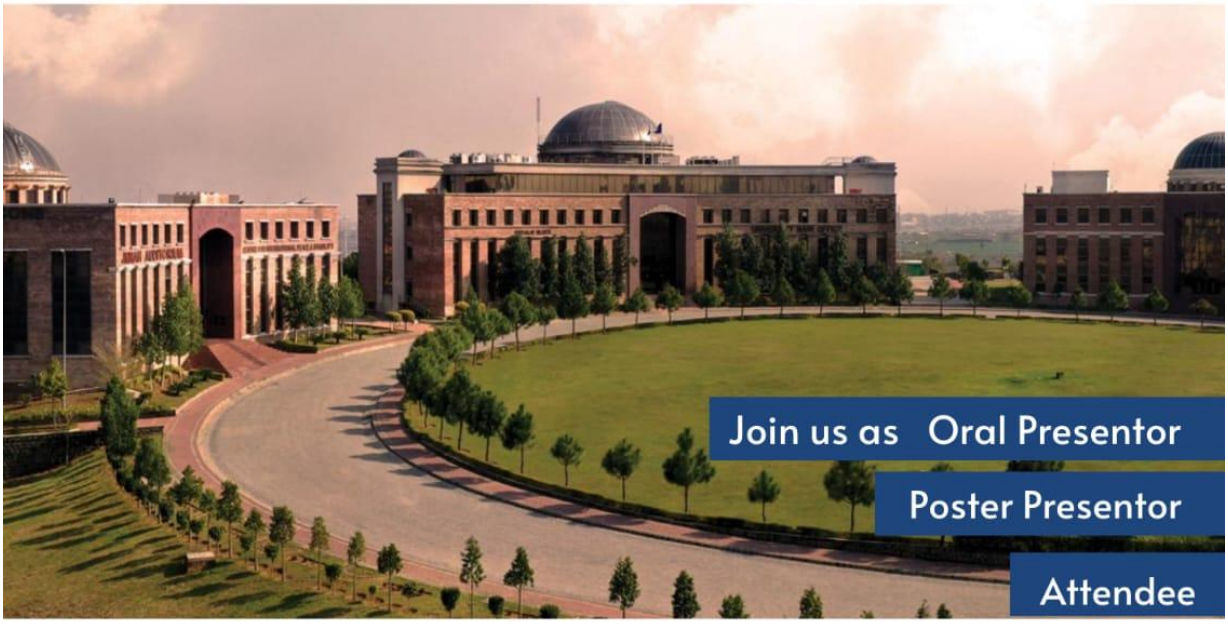
NUST Defining futures School of Natural Sciences

3rd International Conference on Recent Trends in Statistics and Data Analytics

19 -20 December 2024



Dead Line for Registration
28 NOVEMBER 2024

Registration Fee
Faculty/Professional: Rs 3500
Students : Rs 2000

REGISTER NOW

https://forms.gle/WV96BMCkxK6cLdsp8

Contact info:
Dr. Tahir Mehmood 0331-5966869
Dr. Firdos Khan 0345-2433811

Preface

It is with great pleasure that we welcome you to the **3rd International Conference on Recent Trends in Statistics and Data Analytics** organized by the School of Natural Sciences (SNS), National University of Sciences and Technology (NUST), H-12 Sector, Islamabad. This esteemed gathering serves as a platform for researchers, academicians, and industry professionals to exchange knowledge, present innovative ideas, and explore the latest advancements in the field of statistics and data analytics.

In today's data-driven world, statistical methods and data science techniques play a crucial role in addressing complex challenges across various domains, including climate science, economics, healthcare, artificial intelligence, social sciences, engineering, astronomy and natural sciences. This conference brings together experts from diverse backgrounds to discuss novel methodologies, emerging trends, and practical applications that are shaping future data analytics.

The conference features **keynote addresses by distinguished speakers**, **oral and poster presentations**, and **discussions**, covering a wide spectrum of topics, including machine learning, Bayesian statistics, time series analysis, spatial statistics, quantum computing, high dimensional statistical modelling, and big data analytics.

We extend our heartfelt gratitude to all keynote speakers, authors, and participants for their invaluable contributions. Special thanks to our organizing committee and sponsors, whose support has been instrumental in making this event possible. We are eagerly looking forward to the next edition of this event which will be held at the end of 2025.

Dr. Tahir Mehmood	Dr. Firdos Khan	Dr. Mujeeb-Ur-Rehma
Conference Chair	**Conference Focal Person**	**HoD Department of Mathematics**

Table of Contents

Climate Factors as Determinants of Covid-19 Mortality in Pakistan Over the Period of 2020-2023

Muhammad Bilal1*, Dr. Muhammad Mohsin2

1College of Statistical Sciences, University of the Punjab Lahore, Pakistan

2College of Statistical Sciences, University of the Punjab Lahore, Pakistan

*Corresponding Author's Email: bilal7573174@email.com

The COVID-19 pandemic has impacted over 207 countries and territories worldwide, posing significant health and socio-economic challenges. This study explores the influence of climate factors on COVID-19 mortality in Pakistan from 2020 to 2023. Using demographic data on COVID-19 deaths and climate data from March 19, 2020, to March 23, 2023, we performed various statistical analyses, including correlation analysis, principal component analysis, and robust regression. Our findings reveal that climate variables such as all-sky and clear-sky surface shortwave irradiance, surface longwave irradiance, surface PAR (Photosynthetically Active Radiation) totals, UVA and UVB irradiance, temperature, dew point, frost point, and wet bulb temperature significantly influence COVID-19 mortality in Pakistan. These results underscore the importance of considering a wide range of climatic factors, beyond just temperature and humidity, in understanding the pandemic's impact. The insights gained from this research are valuable for global health organizations and local authorities in mitigating COVID-19 deaths and managing future pandemics. This study also provides a deeper understanding of the complex relationship between climate variables and public health outcomes during the pandemic.

Keywords: COVID-19, Pakistan's climate, correlation analysis, factor analysis, robust regression

An efficient MEWMA chart for Gumbel's bivariate Pareto distribution

Ayesha Talib, Sajid Ali and Ismail Shah

1 Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan

2 Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan

3 Department of Statistical Sciences, University of Padua, Padova, Italy

AYESHA.TALIB@hotmail.com

sajidali@qau.edu.pk, sajidali.qau@hotmail.com

Control charts play a vital role in process monitoring to ensure the product's desired standards. Due to rapid improvements in data collection methods, multivariate charts are preferred over univariate charts. This paper proposes a bivariate exponentially weighted moving average chart for the simultaneous monitoring of the mean vector of Gumbel's bivariate Pareto type II (also known as the Lomax distribution) time-between-events data. The performance of the proposed chart is assessed through average run length, median run length, and the standard deviation of the run length criteria. To show the implementation of the chart in the real world, illustrative examples are also presented.

Keywords: Time-between-events; bivariate Gumbel distribution; EWMA chart; average run length

A Novel Robust Adaptive Decomposition Technique for Financial Time-Series

Laiba Sultan Dar, Muhammad Aamir*

1 Statistics, Abdul Wali Khan University Mardan, Pakistan

2 Statistics, Abdul Wali Khan University Mardan, Pakistan

*Corresponding Author's Email: aamirkhan@awkum.edu.pk

Forecasting stock and crude oil prices is essential for ensuring economic stability, enabling strategic decision-making, and managing risks effectively. Accurate predictions support policy formulation, optimize business operations, and enhance investment strategies. Therefore, despite substantial advancements in forecasting, challenges remain in traditional decomposition methods like EMD, EEMD, and CEEMDAN, particularly when there is a high volatility in time-series data. This study introduces Robust Adaptive Decomposition (RAD), a pioneering technique specifically designed for complex time-series data such as Stock and Brent Oil prices. The RAD approach utilizes a multi-step process: (i) assigning adaptive weights to each data point, (ii) calculating weighted means of adjacent values, (iii) constructing smooth cubic splines through these weighted points, (iv) iterative sifting with a novel methodology, and (v) halting based on a precise stopping criterion. The efficacy of RAD was validated by comparing RAD-ARIMA and RAD-LSTM models against six hybrid methods—EMD-ARIMA, EMD-LSTM, EEMD-ARIMA, EEMD-LSTM, CEEMDAN-ARIMA, and CEEMDAN-LSTM. RAD consistently outperformed these established techniques in simulated scenarios and real-world datasets, including daily Brent oil and stock prices. By addressing critical limitations of traditional decomposition methods, such as difficulty in model fitting and labor-intensive processes, RAD sets a new benchmark for time-series analysis, offering a solution for forecasting in volatile and complex financial markets and proposes an improved mode extraction decomposition technique to extract meaningful information from time series data by separating the noise.

Keywords: Robust Adaptive Decomposition Technique (RAD), Empirical Mode Decomposition (EMD), Ensemble Empirical Mode Decomposition (EEMD), Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), Autoregressive integrated moving average (ARIMA), Long short-term memory (LSTM).

Influence Diagnostic Methods for Log-Normal Regression Model

Muhammad Habib, Muhammad Amin

University of Sargodha, Sargodha Pakistan

Department of Statistics, University of Sargodha, Sargodha Pakistan

*Corresponding Author's Email: habibhakim1048@gmail.com

Influential analysis is the primary diagnostic procedure in regression analysis to get efficient and reliable results. In this study, we want to find an effective diagnostic method with appropriate residuals to detect influential point in the log-normal regression model. For this purpose, we propose Cook's distance, modified Cook's distance, Covariance ratio and Hadi methods with fitted and quantile residuals for the log-normal regression. We evaluate the influential point detection capability of these methods with fitted and quantile residuals with help of simulation study and real-world applications. The results indicate that the efficiency of the Cook's distance, CVR and Hadi method with both fitted and quantile residual is superior than the modified Cook's distance method.

Keywords: Influential observation, Log-Normal regression, Cook's distance, Modified Cook's distance, CVR, Hadi method, Fitted residual, Quantile residual

A Functional Data Approach for Forecasting Mean Sea Level Pressure

Muhammad Uzair, Ismail Shah*

Department of Statistics, Quaid-i-Azam University, Islamabad 45320, Pakistan

*Corresponding Author's Email: ismail.shah@unipd.it

The engineering and management of water assets, navigation, and coastal operations all depend on the ability to forecast mean sea level pressure (MSLP) using historical time series data. Based on functional data analysis (FDA), this work suggests a computationally effective method to improve the short-term forecasting precision for MSLP. Splitting cleaned time series data into predictable and probabilistic components is the foundation of this approach. The generalized additive modeling method is used to model and forecast the deterministic component, which includes trends and seasonality. In contrast, functional autoregressive (FAR) and FAR with exogenous variates (FARX) are used to model and forecast the stochastic component. To evaluate the models' performance, one-day-ahead out-of-sample forecasts for a whole year are gathered, together with MSLP data for the Dublin airport site from Ireland's open data portal. The results of a statistical significance test and other accuracy metrics indicate that the suggested functional models outperform their rivals in MSLP forecasting.

Keywords: Mean sea level pressure, Functional autoregressive models, ARIMA, NNAR, Forecasting

Advanced Statistical Techniques for Analyzing Pakistani Children's Health Data: A Comparative Study of GAMLSS and Quantile Regression Methods

Natasha Akbar1, Muhammad Aslam2

1,2*Department of Statistics, Bahauddin Zakariya University, Multan 60800, Pakistan*

Email address: natashaakbar86@gmail.com

Age-related body mass index (BMI) has been suggested by healthcare professionals as an alternative way of assessing childhood obesity. Currently, Balochistan, Pakistan, does not have any established age-related BMI reference curves. This study aims to provide standardized growth reference charts of BMI-for-age, children 1 – 60 months using the generalized additive model for location scale and shape (GAMLSS) and quantile regression (QR) methods. Additionally, each method's applicability in detecting the growth abnormalities was evaluated. The weight and height of 10,297 children including 5,213 (50.6%) boys and 5,084 (49.4%) girls from Balochistan, Pakistan were collected through a cross-sectional study. BMI was calculated as weight in kg divided by height in squared meters. The GAMLSS with box-cox power exponential (BCPE) distribution for non-gaussian data and with the goodness of fit test called worm plots and Q-statistics were used to construct the growth reference charts. To provide flexibility for data that is not normally distributed, QR was also employed as an alternate method for directly estimating conditional percentiles without making any assumptions about distribution. The mean ± (sd) of BMI for both boys and girls were: 16.1 ± (3.4) kg/m2 and 15.8 ± (3.6) kg/m2 respectively. The 3rd, 5th, 15th, 25th, 50th, 75th, 85th, 95th and 97th smoothed percentile values for BMI-for-age for both sexes were presented. Median BMI increased steeply in the early age of months, with a peak at 12 months then declined about 60 months in both sexes. From 1 to 30 months, the boys had greater BMI in contrast with girls. The P50 percentile for Balochistan children was lower than those of comparable international studies for the same age group. In contrast with existing growth references, the median percentile at 3 months was lower in both boys and girls. However, from 3 – 60 months were higher than the national references. Moreover, the median percentile of children living in Balochistan was lower as compared to the WHO 2006 growth standards. The results of this study indicated that the cut-off points for obesity were lower than those of the international average. However, suggesting the nutrition status of Balochistan children is lower than that of children in developed countries, and has not reached the international average level. This study provides the first reference for the children residing in Balochistan from 1 – 60 months. The findings of the current study concluded the possibility of preparation of local growth charts and their importance in evaluating children's growth. Additionally, variations from the WHO standards highlight the need for local charts to be created in any future research involving longitudinal data. Concluded as a comparative analysis we found that different techniques had varying degrees of

success in reflecting the asymmetric and heterogeneous character of BMI distributions among children. Both approaches demonstrated the ability to produce reliable and clinically meaningful growth charts, with GAMLSS excelling in parametric flexibility and QR offering computational simplicity.

Keywords: body mass index, Box-Cox power exponential distribution, growth reference chart, lambda-mu-sigma method, quantile regression, smoothed percentiles.

Negative Binomial Regression Model Estimation using Stein Approach: Simulation and Application

Bushra Ashraf1, Muhammad Amin2*

1 Govt Associate College for Women, Kuthiala Sheikhan Mandi Bahauddin, Pakistan

2 Department of Statistics, University of Sargodha, Sargodha, Pakistan.

*Corresponding Author's Email: muhammad.amin@uos.edu.pk

The Negative Binomial Regression Model (NBRM) is popular for modeling count data and addressing over-dispersion issues. Generally, the Maximum Likelihood Estimator (MLE) is used to estimate the NBRM coefficients. However, when the explanatory variables in the NBRM are correlated, the MLE yields inaccurate estimates. To tackle this challenge, we propose a James-Stein Estimator for the NBRM. The matrix mean squared error (MSE) and the scalar MSE properties are derived and compared with other estimators, including the ridge estimator (RE), Liu estimator (LE) and the MLE. We assess the performance of the suggested estimator using two real applications and a simulation study, with MSE serving as the assessment criterion. Results from both simulations and real applications demonstrate the superior performance of the proposed estimator over the RE, LE, and MLE.

Keywords: Count Model, Multicollinearity, Negative Binomial Regression, Ridge Estimator, Stein Estimator

A Comparative Analysis of Hybrid Machine Learning Techniques for Energy Load Prediction

Warsha Ashraf1*, Roman Zainab2, Muhammad Amin1

1 Department of Statistics, University of Sargodha, Sargodha, Pakistan.

2 Department of Statistics, Bahauddin Zakriya University, Multan-Pakistan

*Corresponding Author's Email: faiqueraza001@gmail.com

Accurate prediction of energy utilization is vital for effective energy preservation, supervision and its strategic organization. Enhancing and refining prediction models is essential for maintaining long-lasting energy systems. In this regard, the current study introduces an innovative hybrid machine learning model designed to forecast the cooling load (CL) and heating load (HL) of residential buildings. This new model, termed group support vector regression (GSVR), combines the group method of data handling (GMDH) and support vector regression (SVR) models to predict the HL and CL. The study also employed foundational methods such as elastic-net regression (ENR), back-propagation neural network (BPNN), k-nearest neighbors (KNN), partial least squares regression (PLSR), general regression neural network (GRNN), GMDH, and SVR. To serve CL and HL as the output variables for each network operational parameters of the buildings were utilized as input variables. After training and initial testing, all models were preserved as black boxes. Lastly, a comparative analysis was conducted to evaluate the predictive performance of the proposed hybrid model against these established methods. The findings of the study showed that proposed hybrid method demonstrated a high coefficient of determination (R^2=99.92%) for CL prediction and 99.99% for HL prediction with minimal statistical errors, and delivering the most accurate prediction performance.

Key words: Heating load, Cooling load, Prediction, Machine learning, Artificial neural network, Regression.

Are Industrial Portfolios Leading Indicators of Sectoral Economic Activity?

Dr. Javed Iqbal,
IBA Karachi

The study investigates whether industry-specific stock portfolios can serve as leading indicators for sector-level economic activity. The research aims to address concerns about stock prices yielding ambiguous signals due to varying business cycle sensitivities and liquidity levels across industries. It is found that for certain industries, such as consumer goods, machinery, and utilities, stock prices improve the predictive ability of associated sectoral industrial production growth. However, for some industries, such as aircraft, automobiles, and textiles, adding stock portfolio returns as predictors worsens the accuracy of long-horizon forecasts compared to autoregressive models. The study employs Granger causality tests and Diebold-Mariano's test of equal predictive accuracy to evaluate the performance of industry-specific stock portfolios as leading indicators. The study contributes to the ongoing debate about the relationship between stock prices and economic activity, providing new insights into the industry-specific nature of this relationship.

Adaptability and Outcomes of Ai-Driven Personalized Learning: A Study Among University Students in Pakistan

Anam Zakir1*

1Department of Statistics, Virtual University of Pakistan, Pakistan

*Corresponding Author's Email: anam.zakir@vu.edu.pk

This study investigates the adaptability and outcomes of AI-driven personalized learning among university students in Pakistan, while also considering its effects on learning accessibility, critical thinking, and self-directed learning. The research used a structured questionnaire targeting 500 students across diverse demographics, assessing their familiarity, usage patterns, and perceptions of AI tools such as ChatGPT and Google Bard. For this study, a technology acceptance model has been considered, which focuses on the perceived ease and usefulness of a technology. Findings indicate that 68% of students agreed with the role of AI in enhancing learning outcomes, 60% agreed that it increased motivation and critical thinking, 72% agreed that it helped individual learning needs and delivered materials at their pace, while 65% improved their writing. However, the challenges of a digital literacy gap and concern for ethical implications were highlighted. In fact, 34% of students expressed fear regarding the ethical implications of AI in education. These results underscore the transforming potential of AI in the context of personalized learning for educators and policymakers to adapt AI integration in higher education inside developing countries.

Keywords: Artificial intelligence, AI-driven learning, accessibility, critical thinking, technology acceptance

Machine Learning-Based Classification of Malnutrition in Children under Five Years in Pakistan: Insights from the Pdhs 2017-18

Anjum Shahzad1*, Tahir Mehmood2

1. SNS, NUST, Islamabad, Kohsar University Murree, Pakistan

2. SNS, NUST, Islamabad, Pakistan

*Corresponding Author's Email: anjum.phdstats23sns@student.nust.edu.pk

Malnutrition significantly impacts newborns under five years age worldwide. Children who consume a balanced diet, including essential nutrients, tend to be healthier compared to those who did not access to adequate nutrition. This study is based upon data from Pakistan's Demographic and Health Survey (PDHS) 2017–2018, conducted every five years in developing countries with the support of UNICEF. The data includes 12707 respondents from all over the Pakistan. There were initially 85 variables included in the survey related to nutritional that may cause the BMI (Body Mass Index) of the children. In the filtering stage of the data, the constant response variable as well as redundant variables and variables have multicollinearity issue have been removed and at the final stage we have left with 46 variables. The response variable was BMI, which converted to latent variable on the basis of gold standard given by World Health Organization (WHO) that a bay have BMI less than -2 z score will be consider under nutrition and a bay have BMI greater than -2 z score will be considered not under under nutrition. The six machine learning techniques have been applied support vector machine (SVM), K nearest neighbors (KNN), logistic classifier, linear discriminant analysis (LDA), Decision Tree (DT) and Naïve Bayes (NB). The logistic regression outperform all other ML techniques for predicting of under nutrition status of under five years babies with highest are under the operating receiving curve (ROC) followed by LDA, Decision Tree, KNN, Naïve Bayes and SVM. The variable selection have been applied by using all six techniques for which again logistic classifier outperforms all other ML techniques with highest area under the curve, i.e. 0.63, followed by KNN, LDA, DT, NB and SVM. The logistic classifier performed significantly better than all other methods for prediction of under nutrition under five years children in Pakistan with highest area under the curve with highest accuracy among all the methods

Keywords: BMI, WHO, PDHS, Logistic Classifier, ROC

A New Proposed Logistic Cotangent Topp-Leone Gumbel Distribution: Properties and Applications

Maryam Siddiqa[1], Huda Uroojh[2]

1,2 Department of Mathematics and Statistics International Islamic University, Islamabad, Pakistan

*Corresponding Author's Email: Maryam.siddiqa@iiui.edu.pk

Lifetime distributions are essential for describing real-world phenomena in various scientific domains. Existing distributions often struggle to capture complex real-world data accurately. One of these distributions that is well-known for its broad applicability in both theoretical and practical fields is the Gumbel distribution. A new modified life time distribution that is more flexible than the current extension of the Gumbel distribution has been proposed in this study. At the first step, a more versatile generator has been introduced to generate a Cotanget Topp-Leone Generalized family of distributions. Secondly, new generator is employed with the base line distribution to formulate a new Gumbel distribution extension with three-parameters. Numerous statistical features of the propose distribution have been presented. The new introduced distribution is contrasted with its existing distributional form. New developed Logistic Cotangent Topp-Leone Gumbel distribution is shown to outperform existing models in real data applications, making it a better option for modeling.

Keywords: Distributions, Trigonometric functions, Lifetime, Topp-Leone

Segmenting and Classifying Skin Cancer Using Sam-Transformer for Automated Analysis an a Noninvasive Digital System

Galib Muhammad Shahriar Himel 1*, Md. Masudul Islam 2, Anusha Achuthan 1,

Kh. Abdullah Al-Aff 3, Shams Ibne Karim 4

1 School of Computer Sciences, Universiti Sains Malaysia, Malaysia.

2 Department of Computer Science and Engineering, Jahangirnagar University, Bangladesh.

3 Department of Microbiology, University of Szeged, Hungary.

4 Department of Psychiatry, Bangabandhu Sheikh Mujib Medical University, Bangladesh.

*Corresponding Author's Email: galib.muhammad.shahriar@gmail.com

Skin cancer remains one of the most pressing health challenges worldwide, necessitating prompt and precise detection to improve survival rates for those affected. The advancements in deep learning techniques have significantly simplified the task of image recognition, making these models increasingly vital in medical diagnostics. This study focuses on the application of a sophisticated deep-learning architecture known as the Vision Transformer for skin cancer detection. The research utilizes the HAM10000 dataset, a publicly accessible collection comprising 10,015 dermatoscopic images of skin lesions categorized into seven subclasses, which include both benign and malignant types. These images are particularly reliable as they are derived from high-quality dermatoscopic examinations conducted by certified dermatologists. To improve the model's robustness and generalizability, various preprocessing techniques such as normalization and data augmentation were employed. The Vision Transformer was further refined and optimized to meet the specific requirements of skin cancer classification. A segmentation model, SAM, was adapted for skin cancer segmentation tasks, yielding impressive results with an IoU of 96%. For classification purposes, several pre-trained Vision Transformer models were evaluated. The experiments demonstrated that Google's ViT model achieved an accuracy of almost 97% with a low false negative rate on the test dataset, highlighting the model's potential to assist dermatologists in diagnosing skin cancer effectively.

Keywords: Skin Cancer Classification, Vision Transformer, Deep Learning, Segmentation, Image Processing, Image Classification.

Probabilistic Approach to Predict the Bowlers for the First Two and Last Two Overs in T-20 International Cricket using Machine Learning Algorithms

Author: [1]Ikram Ullah, and [2]Qamruz Zaman

[1,2]Department of Statistics, University of Peshawar, Pakistan

*Corresponding Author's Email: cricssportsresearchgroup@gmail.com, ikkoblue10@gmail.com

This study aims to create models using artificial neural networks to predict cricket bowler performance during crucial phases of matches, specifically the first two and last two overs. Analyzing data from 167 T-20 international men's cricket bowlers between 2021 and 2023, the research investigated how captains and coaches decide on bowler selection based on metrics like Bowling Strike Rate, Bowling Economy, and Bowling Average. By splitting the data into separate sets for these key overs and using heuristic rules, bowlers were classified as "Selected," "Recommended," or "Dropped" depending on their performance. The study demonstrated the effectiveness of using advanced machine learning techniques in improving decision-making processes and optimizing team compositions in cricket, with high accuracy in predicting bowler classifications observed across multiple iterations of the artificial neural networks. Additionally, the research highlighted the importance of certain independent variables in predicting bowler performance, aiding in refining selection strategies. Overall, the study underscores the potential of advanced analytics in enhancing performance optimization in cricket.

Keywords: Cricket, Predictive Models, Artificial Neural Networks, Bowler Performance, Strategic Decisions, Machine Learning Techniques

Statistical Modeling of the Performance of Unorthodox Players in T-20 International Cricket

Jawad Ullah, Qamruz Zaman

Department of Statistics, University of Peshawar, Pakistan

cricssportsresearchgroup@gmail.com, jdstats5544@gmail.com

This study will specifically focus on the dataset of 25 unorthodox batsmen and 25 unconventional bowlers from 2005 to 2024. Considering that they are selected under the category unconventional players in T-20 international cricket. The information were sourced from Cricbuzz and ESPN Cricinfo. The objectives of the study are as follows. The study aims to: model player performance, classify player types, and identify key predictive factors. A random forest model will be employed to classify the batsmen which resulted in an Out-of-Bag error rate of 20%, with accuracies of 82%, 73%, and 100% for the "Good," "Moderate," and "Poor" categories, respectively. Some other key findings are notably, “Fours”, “Highest Score” and “Fifties” were used as crucial predictors. On the other hand, an 11% error rate was established with the model for bowlers, with 80%, 100%, and 87.5% accuracy rates for the respective performance categories. This is considering "Matches Played" and "Wickets per Match Rate" as significant factors. Although data reliability and availability pose certain challenges, this study emphasizes the critical role that unconventional players play in T-20 cricket. It provides valuable insights that can significantly aid in both talent identification and strategic planning.

Keywords: T-20 cricket, unorthodox players, random forest, player performance, predictive modeling

Early Childhood Malnutrition: An Overview from 26 Low and Lower-Middle Income Countries using Multiple Indicator Cluster Survey 2017-23

Sidra Younas[1]*, Maryam Sadiq[1].

1Department of Statistics, The University of Azad Jammu and Kashmir, Muzaffarabad, Azad Kahmir, Pakistan.

* sidmalik047@gmail.com

Regarding Sustainable Development Goal 3, reduction of Malnutrition is an essential component of good health and well-being of young children. This study added to the body of knowledge about Malnutrition among young children belonging to 26 low and lower-middle income countries. The analysis used the datasets obtained from Multiple Indicator Cluster Surveys conducted between 2017 and 2023, including 323480 children under 5 years of age. The measure of nutritional status consists of three main dimensions: weight-for-age, height-for-age, and weight-for-height to explore wasting, stunting, and underweight among young children. The present analysis aimed to provide an overview of nutritional conditions of young children by examining the distribution of observations regarding income level of family, area and country of residence, mother's education level, number of household members, treating water to make it safer for drinking, availability of toilet facility, gender, education level of household head, and global region. The study also evaluated the country wise descriptive measures for stunting (height-for-age Z-scores), wasting (weight-for-age Z-scores), and underweight (weight-for-height Z-scores). The association of income level of family with the risk of stunting, wasting and underweight among children is represented through Bi-Plots complied by correspondence analysis. The findings of this study significantly provide evidence to enhance the income level of family to improve early childhood nutritional conditions. Remarkable strategies are required to attain Sustainable Development Goals regarding good health at a national and international level.

Keywords: Sustainable Development Goals, Multiple Indicator Cluster Survey, Malnutrition, Stunting, Wasting.

A Visual Analytics System Elevating Data Quality of Time-Series Streaming Data for Machine Learning and Incorporating the Human-In-The-Loop

Syed Muhammad A. Gardezi, Dr. Faisal Cheema, Dr. Ramoza Ahsan

1 Department of Computer Science, National University of Computer and Emerging Sciences (FAST-NU), Islamabad, Pakistan

2 Department of Computer Science, National University of Computer and Emerging Sciences (FAST-NU), Islamabad, Pakistan

3 Department of Data Science and Artificial Intelligence, National University of Computer and Emerging Sciences (FAST-NU), Islamabad, Pakistan

*Corresponding Author's Email: engrmgardezi@gmail.com

The performance of Machine Learning (ML) models in the age of big data largely depends on the quality of the input data, particularly in streaming time-series data, which is continuously changing. However, existing visual analytics systems lack comprehensive solutions for real-time data management, dynamic data imputation, and mechanisms for real-time data drift monitoring, which often results in reduced model performance. To address these challenges, we introduced StreamCure Analytics (SCA) a visual analytics platform that incorporates human feedback and process-driven methods to improve data quality in real-time and perform time series data analytics. The platform enables users to continuously monitor and improve data quality in real-time by providing tools for automated data quality such as detecting missing values, duplicate data, outliers' detection, data drift management and providing dynamic data curation options. It helps users to handle streaming data issues, facilitating visualization with real time adjustments, and make sure that ML models receive good-quality input data for enhanced model performance. An empirical study was conducted across different tasks such as Exploratory Data Analysis (EDA), data preprocessing, machine learning, and streaming data analysis. Results showed high satisfaction, with Cronbach's Alpha between 0.84 and 0.89, indicating strong reliability of our proposed platform. Comparative analysis showed that the StreamCure Analytics system excelled in usability and task speed, meeting 93% of user expectations. Experts in data visualization, big data, and machine learning evaluated the platform, praising its usability and clear visuals. These findings suggest that StreamCure effectively manages real-time data quality, maintaining metrics such as completeness, reliability, and consistency to enhance ML model accuracy and efficiency.

Keywords: Machine Learning Model Performance, Automated Data Quality Assessment, Human-in-the-Loop Analytics, Automated Data Curation, Visual Analytics for Streaming Data, Data Drift Detection

A Machine Learning Framework for Variable Selection in Perinatal Mortality Classification

Ramla Shah*, Dr. Maryam Sadiq

1 Department of Statistics, University of Azad Jammu & Kashmir, Muzaffarabad, Azad Kashmir, Pakistan

2 Department of Statistics, University of Azad Jammu & Kashmir, Muzaffarabad, Azad Kashmir, Pakistan

*Corresponding Author's Email: ramlashah193@gmail.com

A novel variable selection technique termed as CARS-Logistic is proposed by fusing competitive adaptive re-weighted sampling (CARS) and logistic regression. In the current study, the modulus of regression coefficients of the logistic model is utilized for assessing the significance of variables. The capability of CARS-Logistic is assessed using two data sets: real-life and SIMUL datasets. The outcomes disclose that the proposed method is efficient than classical variable selection methods. It can select more significant variables with the least AIC and BIC values together with greater values of three Pseudo R-squared(s). The CARS-Logistic method has selected more significant variables and has lower values of AIC (6349.034) and BIC (6679.074) but greater values of $RM2$(0.062), $Radj\ M2$ (0.047) and $RC\ \&\ S2$ (0.073). The study also points out the factors leading to perinatal mortality in Pakistan. The identified hazards communicate social, cultural, financial, and health-related characteristics of mothers.

Keywords: Machine learning, variable selection algorithm, CARS-Logistic, Perinatal mortality.

Smog in Pakistan | 3rd International Conference on Sustainable Development Goals: Localizing Sdgs Through Academia

Ammara Nawaz Cheema, Muhammad Ahad Khan*, Saadia Khan*, Mahwish Aziz

Dept of Statistics & Data Analytical Sciences, Air University, Pakistan

*Email: mm.ahadkhan384@gmail.com, sadia.khann68@gmail.com

A Call for Action Towards "Cleaner Air" highlights the critical air quality crisis faced by Pakistan, emphasizing its impact on public health, environmental sustainability, and the economy. With Lahore experiencing an Air Quality Index (AQI) exceeding 1100 in November 2024, smog has become a silent killer, exacerbating respiratory diseases, reducing visibility, and harming ecosystems. This poster identifies the primary contributors to smog, including vehicle emissions (40%), industrial pollution (30%), and crop burning (10%), alongside domestic activities. The research underscores the disproportionate effects on outdoor-exposed males and highlights the potential loss of 4-5 years in life expectancy in smog-affected regions such as Punjab and Islamabad. Individual actions, like planting trees (1 tree absorbs 48 lbs of CO_2 annually), promoting public transport, and switching to energy-efficient appliances, are proposed as immediate mitigation steps. Additionally, lessons from countries like China offer a policy-driven pathway to control pollution through emission controls, industrial regulation, and technological innovation. This study calls for urgent action from policymakers, individuals, and stakeholders to implement sustainable solutions and ensure a cleaner, breathable future for Pakistan's population

Keywords: Silent Killer, Grey Blanket, Pollution Pandemic,Airpocalypse in Pakistan,Breathless Cities, Pakistan's Hazy Crisis

Bone Fracture Classification using Deep Learning

Syeda Moazma Fatima*, Tahir Mehmood

SNS, NUST, Islamabad, Pakistan

*Corresponding Author's Email: moazmafatima464@gmail.com

In this article, I will examine the classification of bone fractures, which are critical for protecting vital organs such as the heart and lungs. Bone fractures is a significant challenge in healthcare, as manually analyzing radiological images to locate fractures can be time-consuming and error-prone. Recent research demonstrates that deep learning can effectively perform complex interpretations at the level of healthcare professionals. In this study, I will investigate various architectures, including GoogleNet, Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN), to improve the classification accuracy of different bone fractures. The research will utilize the Bone Fracture Classification dataset from Kaggle, consisting of 989 training images across 10 fracture classes and an additional 140 test images. I will assess the performance of the GoogleNet, ANN, and CNN models using multiple evaluation metrics.

Keywords: Bone fracture, radiological image, CNN, ANN, Kaggle

Classical and Bayesian Estimation of N-Fold Convolution for Power Function Distribution

Farzana Noor

International Islamic University Islamabad, Pakistan

*Corresponding Author's Email: farzana.akhtar@iiu.edu.pk

Power function is one of the important probability distribution and is widely used in different fields like health, lifetime data, biomedical research, electrical electronic, reliability theory, risk analysis and operation research. In this research, we derive n-fold convolution for the power function distribution considering same and different parameters. The derived convoluted power function distribution is then analyzed under Bayesian and classical framework. Exponential and gamma priors are used to derive posterior distribution. To derive Bayes estimators, Squared Error Loss Function (SELF) and Quadratic Loss Function (QLF) are used. In classical estimation, the maximum likelihood estimator and associated mean square error are derived. Bayes estimates and maximum likelihood estimates are obtained by using simulation study and two real life data sets. Results from both real life data sets show that SELF performs better than QLF. Comparison between both priors manifests that the exponential prior is better than gamma prior.

Keywords: Convolution; Power function distribution; Prior distribution; Loss function

Energy Poverty and Respiratory Health: Evidence from 26 Low and Lower-Middle Income Countries

Maryam Sadiq1*, Sidra Younas1

1DEPARTMENT OF STATISTICS

UNIVERSITY OF AZAD JAMMU AND KASHMIR,

MUZAFFARABAD, PAKISTAN

*Corresponding Author's Email: hussainulahmad@gmail.com

Energy poverty has been extensively discussed in the literature as a remarkable framework to address global development challenges including poverty, unhealthiness, inequality, and illiteracy. The one-dimensional and multidimensional energy poverty is significantly associated with the occurrence of acute respiratory infections in young children. This study examined the effect of an upgraded multidimensional energy poverty index on the perspective of acute respiratory infections among children under five years belonging to 26 low and lower-middle income countries. The data from Multiple Indicator Cluster Surveys conducted between 2017 and 2023, encompassing 340703 children belonging to 26 countries under 5 years of age is used for analysis. The Multidimensional Energy Poverty Index consists of five main dimensions with 12 indicator factors. The Partial Least Squares regression is executed to examine the association of acute respiratory infections with multidimensional energy poverty index, income level of family, area of residence, and global region among young children. The present analysis reported that the multidimensional energy poverty, area of residence, income level of family, and geographical region have significant influence on the risk of acute respiratory infections among children. Specifically, the risk of acute respiratory infections increases by 34% due to a unit change in energy poverty among children under five years. Additionally, geographical and regional inequalities substantially affect respiratory health. The findings greatly contribute in developing a remarkable policy for attaining Sustainable Development Goals regarding good health and clean energy by addressing the energy poverty in low and lower-middle income countries.

Keywords: Energy poverty, acute respiratory infections, health

Hybridizing LMD, Arima and Xgboost for Accurate Crude Oil Price Predictions

Jawaria Nasir, Muhammad Aamir

(Times New Roman 12, normal, centered)

[1] *Department of Statistics, Abdul Wali Khan University Mardan, Pakistan*

[21] *Department of Statistics, Abdul Wali Khan University Mardan, Pakistan*

*Corresponding Author's Email: aamirkhan@awkum.edu.pk

Crude oil usage brings about several environmental issues, such as air and water pollution, along with the emission of greenhouse gases that contribute to climate change. Additionally, the reliance on oil has historically created significant political and economic tensions, leading to conflicts, struggles for resource control, and instability. In this study, we propose a novel hybrid approach that uses Local Mean Decomposition (LMD) for breaking down the data and determining contribution coefficients, all while maintaining lower computational complexity. The model incorporates XGBOOST along with ARIMA (Auto Regressive Integrated Moving Average), allowing it to predict each reconstructed segment by leveraging deep dynamic learning features. Since oil market turbulence is highly volatile and unpredictable, this hybrid model is designed to account for random fluctuations and external market pressures. Traditional methods are inadequate for predicting oil prices due to their inability to handle such volatility. However, the hybrid model can manage non-linearity and dynamic changes, offering more accurate forecasting by incorporating multiple techniques that reduce the risk of overfitting and enhance predictive power. This approach is validated using standard evaluation metrics like RMSE, MSE, and MAPE, with publicly available data from WTI (West Texas Intermediate), a widely recognized benchmark in the industry.

Keywords: ARIMA, LMD, XGBOOST, Stochastic and deterministic influences, crude oil prices.

Model-Assisted Small Area Estimation of Health Indicators Using Mixed-Effects Random Forests: Unit-Level Estimation

Muhammad Hamza, Shakeel Ahmed, Youngsoon Kim*

1 Department of Bio & Medical Big Data, Gyeongsang National University, South Korea.

2 School of Natural Sciences, National University of Science and Technology, Islamabad, Pakistan.

3 Department of Information & Statistics and Department of Bio & Medical Big Data, Gyeongsang National University, South Korea.

*Corresponding Author's Email: youngsoonkim@gnu.ac.kr

Small area estimation (SAE) provides reliable estimates for subpopulations or geographic areas, particularly when the direct estimates lack precision because of small sample sizes. This paper illustrates a model-assisted approach using the Mixed-Effects Random Forests (MERF) method for incorporating auxiliary information to increase the precision of population mean. The developed method is applied to produce the small area estimates of stunted children by districts in Pakistan using data from the Pakistan Demographic and Health Survey (DHS) 2017-18. The results indicate that the methods of MERF, especially MERF 3, yield more accurate and reliable estimates of population mean with lower variability and reduced bias, even in districts with small sample sizes. This study identifies MERF as a potentially robust model-assisted technique for addressing challenges in small area estimation and informs policymakers with valuable insights for resource allocation and targeted interventions to combat stunting and related public health challenges.

Keywords: Model-assisted estimation, Small area estimation, Random forests, Mixed-effects regression tree, Mixed-effects random forests, Survey data, Stunting

Comparison of Cross Validation and Bayesian Approach for Selection of Optimal Hyper Parameter Selection for Ridge Regression.

1st International Conference on Sustainable Development Goals: Localizing SDGs Through Academia.

NAQASH Shabbir, Ismail Shah Afridi*, Sajid Ali

Department of Statistics Quaid I Azam University Islamabad Pakistan.

Corresponding Author's Email: nshabbir@stat.qau.edu.pk

Abstract

Machine learning techniques rely on internal parameters, known as hyperparameters, which require optimization for optimal performance. This optimization presents a difficulty for machine learning practitioners, either specialized skill, intuition, or resource-intensive parameter search. This study introduces a Bayesian method for calculating the ridge parameter λ in ridge regression. We use ridge parameter (k) values suggested by several authors as prior distribution, from different existing re searches. Using Bayesian estimation, we get the posterior distribution and choose an ideal λ that minimizes the mean squared error (MSE) of the regression coefficients. Simulation results indicate that the Bayesian estimator outperforms cross validation (CV) approaches regarding Bias, Variance and MSE. Our findings indicate that the Bayesian technique offers a more resilient alternative to cross-validation, especially in the context of multicollinearity. This technique can be used to enhance model stability and predictive adequacy in regression analyses.

Keyword: Ridge regression, Cross Validation, Bayesian estimation, Mean square error

Audio Analysis for Urban Sound Classification: A Six-Feature Approach using LSTM and GRU

MOMNA HASSNAIN

SCHOOL OF NATURAL SCIENCES , NUST

Email: momnabiya762@gmail.com

Environmental sound classification has gained significant attention due to its broad applications in urban planning, public safety, and environmental monitoring. However, the unstructured and diverse nature of environmental sounds poses challenges for effective classification. In this study, we address these challenges by leveraging the strong spectro-temporal patterns inherent in audio data through deep learning models specifically designed for sequence modeling: Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. Initially, the models were trained on magnitude Mel-spectrograms and simple audio features, including Short-Time Fourier Transform (STFT), Mel-frequency cepstral coefficients (MFCC), zero-crossing rate (ZCR), chroma features, and constant-Q transform (CQT). These features provided a solid foundation for classification and achieved an accuracy of up to 95\%. To further enhance the model's performance, we introduced advanced techniques such as wavelet transforms as a replacement for the Fourier transform and employed different windowing methods to optimize the feature extraction process. These improvements allowed the models to capture more nuanced patterns in the audio data, ultimately increasing the classification accuracy to 97\%. This research demonstrates the effectiveness of incorporating advanced audio processing techniques with LSTM and GRU architectures for environmental sound classification. The findings underscore the importance of spectral analysis and feature refinement in achieving high accuracy for audio classification tasks, highlighting the potential of deep learning models in this domain.

A Restricted Modified Ridge Type Estimator

1st Roman Zainab*, 2nd Muhammad Aslam*

1st Department of Statistics, Bahauddin Zakariya University, Multan-Pakistan

Email: romanzainab@gmail.com

2nd Department of Statistics, Bahauddin Zakariya University, Multan-Pakistan

Email: aslamasadi@bzu.edu.pk

The linear regression model is a foundational tool in statistical analysis, widely used across academic fields and practical applications. Among the techniques to estimate linear regression models, the ordinary least squares (OLS) estimator is the most commonly employed due to its simplicity and interpretability. Nonetheless, this method's stability wanes, and its results turn misleading when confronted with multicollinearity. This study focuses on multicollinearity issues in a linear model with linear constraints on the parameter vector. To address these challenges, we propose a novel estimator that integrates the restricted least squares (RLS) estimator with a modified ridge-type estimator. Our analysis reveals that this new approach outperforms both the RLS and restricted ridge regression estimators, as shown through a rigorous assessment using the mean squared error criterion, which effectively mitigates the adverse effects of multicollinearity. To validate our approach, we conduct both simulation studies and real-data applications, specifically using the Portland Cement data set. The alignment of our empirical findings with theoretical predictions underscores the estimator's effectiveness and applicability in multicollinear settings.

Keywords: Mean Squared Error; Modified Ridge Type Estimator; Multicollinearity; Restricted Least Square Estimator; Restricted Ridge Regression Estimator

Mobile Price Prediction Without Feature Reduction

Muqaddas Rizwan

FAST National University of Pakistan

muqadascreator@gmail.com

The most fascinating thing that is being used in the 21st century is mobile phones. With the advancements of hardware and software, new mobile versions and new mobile companies are coming in market. To launch a mobile in market, deciding its price is one of the most important questions for a company that needs to be answered. The price of a mobile phone is a crucial factor in determining its success in market. This paper aims to explore the machine learning models that best predicts the mobile price based on its features. The four Machine learning models Logistic Regression, Random Forest Gradient Boosting Classifier, Support Vector Machine have been implemented on open dataset to train and test the accuracy. Our Two implemented Machine learning models have achieved more accuracy than existing models and those are Logistic Regression and Support Vector Machine. Logistic Regression has achieved 95% accuracy and Support Vector Machine has achieved 94% accuracy using open dataset for mobile price prediction.

Keywords: Mobile price prediction using Machine learning, Mobile price prediction, price prediction, Classification

Machine Learning-Based Forecasting of Chlorophyll-A Concentrations using Satellite Remote Sensing Data in the Arabian Sea

Muhammad Tahir and Yao Shutao

School of Space Science and Technology, Shandong University, Weihai, China

Corresponding author: yaoshutao2008@sdu.edu.cn

Forecasting chlorophyll-a (Chl-a) concentrations is crucial for managing marine ecosystems sustainably. It aids in the early detection of environmental changes, such as algal blooms and ecosystem degradation, which impact biodiversity, fisheries, and coastal communities. However, despite its critical importance, most previous research has concentrated on reconstructing and interpreting historical Chl-a data, with limited focus on developing robust and accurate predictive models. This study aims to address this gap by leveraging machine learning and statistical approaches to forecast Chl-a concentrations in the Arabian Sea using satellite-derived time series data. In addition to Chl-a, the analysis incorporates key environmental variables such as sea surface temperature (SST), sea surface height (SSH), and mixed layer depth (MLD), which are known to influence marine productivity. Five predictive models—XGBoost, LSTM, Random Forest, multi-layer perceptron (MLP) and support vector regression (SVR) were evaluated to determine their ability to capture complex temporal dependencies and nonlinear interactions in the data. Model performance was assessed using robust statistical metrics, including root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2). Among the tested models, XGBoost outperformed the others, achieving the highest accuracy with an RMSE of 0.007, R^2 of 0.84, and MAPE of 7.56%. LSTM and MLP models showed competitive performance, while SVR was less effective due to its inability to capture the nonlinear patterns inherent in the data. Residual analysis and feature importance rankings reinforced the robustness and interpretability of the XGBoost model, highlighting its capability to handle the complex dynamics of environmental time series. This study underscores the importance of integrating advanced machine learning techniques using satellite remote sensing data for accurate and reliable ecological forecasting, offering a scalable framework to support proactive marine ecosystem monitoring and sustainable resource management.

Growth and Development of PhD Statistics in Pakistan: Insight from HEC Record

Javed Iqbal
Department of Statistics, Virtual University of Pakistan, Pakistan
Corresponding author Email: *javediqbal@vu.edu.pk, iamjavediqbal@gmail.com*

The evolution of the field of Statistics in Pakistan has been shaped by the growing demand for data-driven decision-making and the emergence of interdisciplinary applications. This study investigates the growth and development of PhD programs in Statistics across Pakistan using data from the Higher Education Commission (HEC) PhD Country Directory, spanning from 2000 to 2023. Key objectives include analyzing trends in PhD dissertations, examining gender and regional disparities, evaluating the impact of supervisors, and identifying dominant research subdomains. The findings reveal a significant gender disparity, with male PhD holders constituting 64% of graduates and an overwhelming 99% of supervisors. Punjab emerged as the leading province in producing PhD holders, accounting for 137 out of 178 graduates, while Baluchistan reported none. Institutions such as the National College of Business Administration & Economics Lahore and Bahauddin Zakariya University Multan were the top contributors to PhD graduations. Probability Distributions, Regression, and Sampling were the most researched subdomains, whereas Demography and Inference had minimal representation. The study underscores critical gaps, including the underrepresentation of female supervisors and regional imbalances, while highlighting the contributions of top universities and supervisors. These insights are intended to inform policy development, promote equity, and advance PhD education in Statistics in Pakistan.

Keywords: PhD Statistics, HEC, PCD, Equity in Education, Data Science

Building a Hybrid Model to Enhance Forecasting Accuracy

Kifayat Ullah[1*], Ijaz Hussain[2], Muhammad Aslam[1], and Mohsin Abbas[3]

[1]Department of Mathematics and Statistics, Institute of Business Management, Karachi, Pakistan.
[2]Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan.
[3]Department of Management Science and Engineering, University of Science and Technology, Beijing, China.

Streamflow plays a vital role in the effective management and allocation of water resources, as well as in power generation systems. However, its inherently linear and nonlinear characteristics make the development of accurate predictive models challenging. Real-world streamflow data often exhibit both linear and nonlinear components, which cannot be effectively modeled using statistical or machine learning methods alone. To address this issue, a hybrid two-stage model is proposed. In the first stage, the linear component is estimated using the ARIMA (Auto-Regressive Integrated Moving Average) model, leaving behind residuals that represent the nonlinear relationships. In the second stage, these residuals are modeled using machine learning techniques. Specifically, we employ hybrid models that combine ARIMA with Neural Network Auto-Regressive (NNAR) and Multi-Layer Perceptron (MLP) approaches. Both MLP and NNAR are well-suited for capturing nonlinear patterns in time series, complementing ARIMA's strength in modeling linear dynamics. The performance of the proposed ARIMA-NNAR and ARIMA-MLP models was evaluated using two datasets from the Tarbela and Jehlum rivers, spanning the period from January 2017 to April 2021. Comparative analyses demonstrated that these hybrid models outperformed standalone models based on key performance metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The results highlight the significant improvement in forecasting accuracy achieved by integrating ARIMA with NNAR and MLP.

Keywords: Time series, ARIMA, Multi-Layer Perceptron, Neural Network Auto-Regressive, Hybrid models, ARIMA-NNAR, ARIMA-MLP

Machine Learning for Early Alzheimer's Disease Prediction on High-Dimensional Multimodal Data

Aunsia Khan1*, Anusha Achuthan2

1School of Computer Sciences, University Sains Malaysia, Malaysia

*aunsia@student.usm.my

T Alzheimer's disease (AD) poses a significant challenge to global healthcare and its prevalence is expected to increase substantially reaching up to 139 million people by 2050. Early prediction is essential for proper medication and treatment. Various research has been conducted to predict AD at earlier stage using different modalities while this research will be based on high dimensional multimodal data comprising of clinical and genetic features. High dimensional multimodal dataset enables machine learning models to perform better by capturing complex patterns in the data. It allows models to learn more comprehensive features leading to more accurate predictions. However, the high dimensionality and large size of the multimodal data pose substantial hurdles in Alzheimer's disease analysis. Recent advancement in Machine learning shows promise in improving predictive capabilities; however, training machine learning models on high-dimensional multimodal datasets remains a costly and complex task that requires careful consideration. These datasets increase the computational requirements such as processing power, memory and significantly increase the model's training time. This study explores the use of machine learning methodologies for the early prediction of Alzheimer's disease while addressing the computational challenges associated with managing the high dimensional multimodal data. The datasets comprise of diverse modalities having information such as clinical assessments which are rich in longitudinal information and genetic data. The large size and high dimensionality of clinical and genetic data requires robust techniques to extract meaningful information for accurate predictions. Traditional methods often struggle to keep up with the growing AD dataset, leading to bottlenecks in data processing and analysis. Therefore, there is a need for optimized machine learning algorithms that can efficiently handle the intricacies of clinical and genetic features based multimodal data. In this research, the importance of feature selection and dimensionality reduction are focused to enhance computational efficiency and prediction accuracy.

Keywords: Multimodal data, Alzheimer's disease, Dimensionality reduction

Are Industrial Portfolios Leading Indicators of Sectoral Economic Activity?

Javed Iqbal

IBA Karachi

Email: jiqbal@iba.edu.pk

The study investigates whether industry-specific stock portfolios can serve as leading indicators for sector-level economic activity. The paper aims to address concerns about stock prices yielding ambiguous signals due to varying business cycle sensitivities and liquidity levels across industries. It is found that for certain industries, such as consumer goods, machinery, and utilities, including industry stock returns improve the predictive ability of associated sectoral industrial production growth. However, for some industries, such as aircraft, automobiles, and textiles, adding stock portfolio returns as predictors worsens the accuracy of long-horizon forecasts compared to autoregressive models. The study employs Granger causality tests and Diebold-Mariano's test of equal predictive accuracy to evaluate the performance of industry-specific stock portfolios as leading indicators. The study contributes to the ongoing debate about the relationship between stock prices and economic activity, providing new insights into the industry-specific nature of this relationship.

Keywords: leading indicator; industrial production growth; industrial portfolios; Granger causality; rolling window forecast

L-Moments Based Variance Estimators using Calibration Approach

Usman Shahzad

Department of Management Science, College of Business Administration, Hunan University, Changsha 410082, China

*Corresponding Author's Email: usman.stat@yahoo.com

This research proposes a new method of variance estimation based on L-moments and calibration methods for StRS and StACS. Thus, L-moments that are less sensitive to the influence of outliers than standard ones provide more accurate estimators of variance in heterogeneous and clustered populations. The method utilizes calibration constraints to reweight samples on the basis of auxiliary data leading to increased efficiency. Finally, the efficiency of the proposed estimators in the presence of outliers is ascertained from numerical simulations and random data patterns of wildlife and artificial sets compared with the conventional approaches. Therefore, this research demonstrates the flexibility of L-moments and calibration for making statistically valid inferences while it directs future research toward the integration of multivariate auxiliary information and different sampling techniques.

Keywords: Variance Estimation, L-moments, StRS, StACS, Calibration

Robust Feature Selection using Margin-Weighted Discrminant Analysis on High Dimensional Imbalanced Gene Expression Data

Sheema Gul, Dost Muhammad Khan*, Zardad Khan

1 Department of Statistics, Abdul Wali Khan University, Mardan, Pakistan

2 Department of Statistics, Abdul Wali Khan University, Mardan, Pakistan

3 Department of Statistics and Business Analytics, United Arab Emirates University, Al Ain, UAE

* dostmuhammad@awkum.edu.pk

High dimensional gene expression data, presents considerable challenges for binary classification relevant to feature selection methods. These methods face challenges to efficiently depict the minority class, resulting in sub-optimal results. To mitigate these issues, customized feature selection methods are required to tackle the class imbalance issue. This study aims to propose a more robust solution for feature selection concerning the aspect of high dimensional imbalanced problems, known as Margin Weighted Robust Discriminant Score (MW-RDS). The score, RDS is weighted by Margin weights extracted from support vector machine (SVM) to enhanced discriminative power of genes/features thereby highlighting its potential for class separation. Finally, top ranked genes are constrained using l_1-regularization to discard redundant genes while identifying the most significant genes set. To assess the performance of the proposed method is tested on 9 openly accessible gene expression datasets, using classifiers in term of performance metrics, consistently proven that MW-RDS outperformed existing methods. Overall, the method showed promising results highlight its ability to improve classification performance while guaranteeing addressing minority class problem. Bubplots and metric-plot are also generated to gain a deeper understanding of the results. The results reveal that MW-RDS surpasses the existing feature selection methods based on performance metrics.

Keywords: Margin Weighted Robust Discriminant Score (MW-RDS)

Multimodal Analysis of Beauty and Diversity on Instagram: A Deep Learning Approach)

Huma Nisar, Farwa Zainab, Marrium Jamil, Hassan Jabbar, Amina Asghar, Mian Ahad Husaain

Department of Computer Science, Government College University Faisalabad, Sahiwal (Campus), Pakistan

humajamil200@gmail.com

This research examines the way beauty and diversity are presented on Instagram with a more technologically potent view of robustly being based on methods that are computationally intensive. High precision advanced image-processing models on CLIP, ViT, Swin Transformer, ResNet, SEER are used to discern, through a quantitative appraisal of visual content, what form of beauty standards have managed to be communicated; also, text-based information culled with the use of the BERT model forms a complement to that very visual analysis of a great play between visuals and the narratives. Then, it moves on to multimodal classification, in which it shows the effectiveness of the LLaVA model to consider visual and textual data and thus make the complex classification task feasible at a high accuracy level. This shows that prompt engineering is a very important factor for the refinement of the model output, and through this, the importance of prompt engineering in multimodal analysis comes to the forefront. Therefore, the Late Fusion approach which combines ViT and BERT sums up to providing a holistic framework that will assist in enriching one's knowledge about beauty and diversity in social media narratives. The systematic approach underscores how integrating various modalities into a system will help tackle the complexities of social media content. Key contributions include the novel application of state-of-the-art deep learning models as well as the development of a multimodal analytic framework that collectively enhance the understanding of subjective constructs in the social media beauty discourse. However, inherent challenges, such as subjective classification and dataset limitations, demand avenues for further work, such as better-quality dataset creation and novel forms of prompting to advance the developing field.

Keywords: beauty, diversity, multimodal, classification, deep learning

Optimizing Copula Regression Models with Elastic Net Regularization

Huma Rani [1*], Tahir Mehmood[2], and Muhammad Aslam [3]

[1] *Department of Basic Sciences, Riphah International University, Islamabad, Pakistan*

[2]*School of Natural Sciences, National University of Sciences and Technology (NUST), Islamabad, Pakistan*

[3]*Department of Basic Sciences, Riphah International University, Islamabad, Pakistan*

**e-mail: huma_waleed2014@hotmail.com*

This research delineates an innovative methodology that integrates the Elastic Net regularization technique for feature selection within the copula regression framework to enhance model accuracy and variable selection in high-dimensional contexts. Copula models are highly regarded for their flexibility in capturing complex interdependencies among variables. Simultaneously, the Elastic Net synergizes the advantages of Lasso (L1) and Ridge (L2) regularization, thereby enabling the simultaneous selection of correlated predictors alongside robust estimation. The proposed Elastic Net copula regression framework is implemented in real-life data. Comparative analysis demonstrates that the Elastic Net copula model not only surpasses conventional copula models in terms of predictive accuracy but also exhibits a greater proficiency in identifying significant predictors. The empirical findings underscore the practical relevance of this methodology in addressing intricate, multidimensional data, thereby offering a comprehensive solution for dependency modelling and feature selection in real-world scenarios.

Keywords: Copula Regression Models, Feature Selection, Elastic Net Regularization,

Regression to the Mean for Overdispersed Count Data

Kiran Iftikhar1, Manzoor Khan2*, Jake Olivier3

1Department of Mathematics and Statistics, University of Agriculture, Faisalabad, Pakistan

2 a Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan

3 School of Mathematics and Statistics, University of New South Wales, Sydney, Australia

kiran.iftikhar@uaf.edu.pk

In repeated measurements, regression to the mean (RTM) is a tendency of subjects with observed extreme values to move closer to the mean when measured a second time. Not accounting for RTM could lead to incorrect decisions such as when observed natural variation is incorrectly attributed to the effect of a treatment/intervention. A strategy for addressing RTM is to decompose the total effect, the expected difference in paired random variables conditional on the first being in the tail of its distribution, into regression to the mean and unbiased treatment effects. The unbiased treatment effect can then be estimated by subtraction. Formulae are available in the literature to quantify RTM for Poisson distributed data which are constrained by mean–variance equivalence, although there are many real-life examples of overdispersed count data that are not well approximated by the Poisson. The negative binomial can be considered an explicit overdispersed Poisson process where the Poisson intensity is chosen from a gamma distribution. In this study, the truncated bivariate negative binomial distribution is used to decompose the total effect formulae into RTM and treatment effects. Maximum likelihood estimators (MLE) and method of moments estimators are developed for the total, RTM, and treatment effects. A simulation study is carried out to investigate the properties of the estimators and compare them with those developed under the assumption of the Poisson process. Data on the incidence of dengue cases reported from 2007 to 2017 are used to estimate the total, RTM, and treatment effects.

Keywords: Bivariate negative binomial distribution, Regression to the mean, Over-dispersion, Treatment effect

Data-Driven Pharmaceutical Pricing: Leveraging Machine Learning for Price Prediction in Developing Markets

1.Abdul Ahad Hassan Farooqi 2. Zahoor Ahmad

Department of Statistics, University of Sargodha, Sargodha, Pakistan

1.abdulahadfarooqi73@gmail.com 2. **zahoor.ahmad@uos.edu.pk**

Pharmaceutical pricing in developing markets is challenged by issues of transparency, fairness, and efficiency. The lack of clear pricing structures hampers informed decision-making among pharmaceutical companies, healthcare providers, and regulators. This study addresses these challenges by using machine learning (ML) techniques to predict pharmaceutical prices based on features like company, pack size, discount, and availability. The objective is to create an accurate model to optimize pricing strategies and enhance market transparency. The research employs two advanced ensemble techniques, stacking and blending, which combine predictions from multiple base models to boost performance. The base models used include XGBoost, Random Forest, Linear Regression, and Feedforward Neural Networks (FNN). The stacking ensemble aggregates predictions using a Linear Regression meta-model, while the blending approach uses a similar meta-model to improve accuracy and generalizability. Performance was evaluated using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE). Results showed that although models like XGBoost performed well individually, the blending ensemble technique delivered the best predictive results and improved generalizability. This underscores the value of ensemble methods, especially blending, in complex pricing tasks. The study offers a data-driven framework for pharmaceutical companies and policymakers to set optimal prices, reduce discrepancies, and promote market transparency in developing countries.

Keywords: Pharmaceutical pricing, Machine learning, Feedforward Neural Networks (FNN), Stacking ensemble, Blending ensemble, XGBoost, Random Forest, Meta-model.

Climate Factors as Determinants of Covid-19 Mortality in Pakistan over the Period of 2020-2023

Muhammad Bilal1*, Dr. Muhammad Mohsin2

1College of Statistical Sciences, University of the Punjab Lahore, Pakistan

2College of Statistical Sciences, University of the Punjab Lahore, Pakistan

*Corresponding Author's Email: bilal7573174@email.com

The COVID-19 pandemic has impacted over 207 countries and territories worldwide, posing significant health and socio-economic challenges. This study explores the influence of climate factors on COVID-19 mortality in Pakistan from 2020 to 2024. Using demographic data on COVID-19 deaths and climate data from March 19, 2020, to March 23, 2023, we performed various statistical analyses, including correlation analysis, principal component analysis, and robust regression. Our findings reveal that climate variables such as all-sky and clear-sky surface shortwave irradiance, surface longwave irradiance, surface PAR (Photosynthetically Active Radiation) totals, UVA and UVB irradiance, temperature, dew point, frost point, and wet bulb temperature significantly influence COVID-19 mortality in Pakistan. These results underscore the importance of considering a wide range of climatic factors, beyond just temperature and humidity, in understanding the pandemic's impact. The insights gained from this research are valuable for global health organizations and local authorities in mitigating COVID-19 deaths and managing future pandemics. This study also provides a deeper understanding of the complex relationship between climate variables and public health outcomes during the pandemic.

Keywords: COVID-19, Pakistan's climate, correlation analysis, factor analysis, robust regression

1st International Conference on "Managing Political Instability: The Impact of BOD Characteristics, Corporate Governance Mechanisms, and Underwriter Reputation on IPO Performance for Enhancing the Industrial Management System"

Muhammad Raza Zafar*, Muhammad Hasnain Ali

Institute of Banking and Finance, Bahauddin Zakariya University Multan, Pakistan

Email: musapmu2009@hotmail.com

In the context of the agency theory and the signaling theory, the purpose of this research is to eradicate information asymmetry by developing corporate governance processes and board of director characteristics that may serve as an indication of business excellence. The presence of a properly formed board of directors during the initial public offering (IPO) process may serve as an indication to prospective investors that the level of quality of the enterprise is of a superior standard. Furthermore, scholars examine the moderating impact of political instability. The research sample encompasses the entire set of 292 primary public offerings that have been initiated subsequent to the inception of the Pakistan stock exchange. The study's sample consisted of 82 organizations. The process of selecting participants was conducted through the utilization of the convenience sampling technique. The time frame scope of our investigation is limited to the period spanning from 2000 to 2022. The findings of current research indicated that the agency theory's board mechanisms can minimize IPO performance by ensuring the accuracy of financial disclosure in prospectuses, minimizing present uncertainty, and reducing information asymmetry. Furthermore, underwriting reputation are positively associated with IPO performance, while BOD composition is negatively associated with IPO performance. This view is strengthened by the signaling theory, which states that companies send signals to eliminate information asymmetries.

Keywords: Industrial Management System, Initial Public Offerings, Corporate Governance Mechanism, Underwriter Reputation, Political Instability

SVM-Based Classification of Microarrays Gene Expression Data

Kashmala Akhtar

School of Natural Sciences, NUST

Classifying microarray gene expression data is crucial due to its high-dimensional nature and its significant impact on disease diagnosis and personalized treatment strategies. Timely and accurate classification of gene expression data greatly influences treatment outcomes and patient survival rates. Traditionally, gene expression data analysis involves various statistical methods. However, with the emergence of advanced machine learning techniques, automated classification within these datasets becomes crucial. Present methodology typically involve SVM classifier with different kernel functions to classify diverse gene expression profiles. Nonetheless, the varied characteristics within gene expression data present notable classification challenges.In our study, we introduce a comprehensive dataset comprising thousands of gene expression profiles from Leukemia cancer. Our approach involves proposing an optimal classification method by fine-tuning Support Vector Machine (SVM) parameters and selecting the most appropriate kernel functions. We utilize both standard and refined SVMs with various kernel functions, including linear, polynomial, radial basis function (RBF), and sigmoid, alongside penalized SVM models using L1, Smoothly Clipped Absolute Deviation (SCAD), and SCAD + L2 penalties to improve classification performance.Notably, our innovative approach, when applied to refined SVM with linear and polynomial kernels, achieves superior performance, with the L1 norm exhibiting the best classification accuracy among penalized models. This breakthrough marks a significant advancement in gene expression data classification literature, highlighting the potential of SVMs, particularly with linear and polynomial kernels combined with appropriate penalty terms, for precise and efficient disease classificatio. boaaccuracy of financial disclosure in prospectuses, minimizing present uncertainty, and reducing information asymmetry. Furthermore, underwriting reputation are positively associated with IPO performance, while BOD composition is negatively associated with IPO performance. This view is strengthened by the signaling theory, which states that companies send signals to eliminate information asymmetries.

Keywords: Support vector machine, microarrays gene expression, kernels, penalities, Leukemia cancer

Predicting the Role of Key Players and Team Formation for T-20 Cricket through Network Analysis

Muhammad Irshad, Qamruz Zaman

Department of Statistics, University of Peshawar

cricsportsresearchgroup@gmail.com

This paper uses network analysis to study how T-20 cricket teams are organized, focusing on the partnerships between batsmen. By mapping out these partnerships, the study explores which players are most connected and how they impact their teams. Surprisingly, it finds that players with the best batting averages or main roles aren't always the ones most central to team dynamics. Additionally, clusters of players from similar eras show how team connections evolve over time. The analysis also reveals that teams like Austria, England, India, and Sri Lanka rely more on top-order batsmen, while Pakistan and the West Indies depend on middle-order players. Players with strong connections in the network can significantly influence team performance, making this analysis useful for refining team strategies and improving player selection in T-20 cricket.

Exploring the Impact of Urbanization and Economic Growth on Environmental Degradation in South Asia: A Bayesian Panel Approach

Qasim shah, Syed Muhammad Asim, Alamgir Khalil

Department of statistics, University of Peshawar, Pakistan.

Email: Qasimshah707@gmail.com

This paper investigates the links between urbanization, economic growth, and environmental degradation in South Asia from 2000 to 2020, focusing on key indicators like CO2 emissions, deforestation, and air quality. Using Bayesian panel estimation with fixed and random effects models, the study offers more robust insights compared to traditional methods by incorporating inverse gamma priors. The findings reveal that GDP per capita, trade openness, foreign direct investment, and urbanization significantly contribute to environmental degradation, with varying impacts across countries. The research highlights the role of urbanization and economic growth in driving pollution and resource depletion while underscoring the need for region-specific strategies. Sustainable urban planning, green infrastructure, and eco-friendly trade policies are crucial for mitigating these effects. This study provides policymakers with evidence-based recommendations to balance economic growth with environmental sustainability, emphasizing the critical need for development approaches that prioritize environmental preservation.

Keywords: Urbanization. Economic Growth. Environmental Degradation. Bayesian Panel Estimation. CO2 Emissions.

Robust Nonparametric EWMA Control Chart using Wilcoxson Signed Rank Test

Rizwan Munir[1] & Muhammad Abid[2*]

[1]*The University of Faisalabad, Pakistan, 38000.*

[2]*Government College University, Faisalabad, Pakistan, 38000.*

**Corresponding Author's e-mail: m.abid@gcuf.edu.pk*

Control charts may aid in maintaining and improving the efficiency of manufacturing and industrial processes. Nonparametric (NP) control charts are more dependable and practical than parametric charts when it is unclear how the data will be distributed. When the distribution of the underlying process is unknown or uncertain, NP control charts are required. The NP charts are a reliable alternative that also have the capacity to quickly detect shifts in process parameter(s). For effective process location monitoring, we have developed a nonparametric extended exponentially weighted moving average chart based on the Wilcoxon signed rank (WSR) test (hereafter named $EEWMA_{WSR}$;). The Proposed $EEWMA_{WSR}$; for in-control (IC) and out-of-control (OOC) processes is computed in the study. The performance of the developed schemes has been evaluated by calculating the run-length (RL) properties using the Monte Carlo simulation approach. This study examines the (IC) behavior and (OOC) efficacy of the suggested chart using normal and non-normal distributions. This study includes the parametric such as the EEWMA, EWMA chart and nonparametric charts, $EEWMA_{WSR}$, $EWMA_{WSR}$ control charts, for the comparative analysis. The proposed chart's practical implementation is also illustrated using a real-life application.

Keywords: Control Charts; $EEWMA_{WSR}$; Parametric Chart, Non-Parametric, Run Length;

Analysis of Hereditary Influences on T-20 International Cricket

Zakir Hussain, Qamruz Zaman

Department of Statistics, University of Peshawar

*Corresponding Author's Email: cricssportsresearchgroup@gmail.com, zekikhan999@uop.edu.pk

In the fascinating realm of T-20 international cricket, familial bonds significantly influence player representation across the top 11 teams India, England, Pakistan, South Africa, New Zealand, Australia, West Indies, Sri Lanka, Bangladesh, Afghanistan, and Zimbabwe. An in-depth analysis reveals that approximately 28.9% of players have familial ties, with 65 pairs of brothers, 38 pairs of fathers and sons, and 16 pairs of cousins among them. These connections extend to uncles and nephews, brothers-in-law, and even unique pairings such as a grandfather and granddaughter. While the distribution of familial ties varies across countries, statistical tests, including t-tests and ANOVA, suggest that these hereditary connections do not necessarily correlate with superior performance. Detailed comparisons of players, such as Pakistani brothers Umar and Kamran Akmal, Cousins Babar Azam, Kamran Akmal, and Umar Akmal, and New Zealand's Kane Williamson and Dane Cleaver, illustrate the diversity in performance outcomes, emphasizing that individual talent and other factors play a crucial role in cricket success.

Keywords: T-20 international cricket, familial bonds, hereditary influence, player performance, statistical analysis,

Analyzing Survival Time Upper Record Values with Inverse Weibull Distribution: A Bayesian Approach

Rabia Azeem[1], Muhammad Aslam[1], Tahir Mehmood[2]

[1]Department of Mathematics and Statistics, Riphah International University Pakistan.

[2]School of Natural Sciences, National University of Science and Technology, Pakistan.

Abstract

The highlights is on the Bayesian and classical estimation of the inverse Weibull distribution for upper record values. Independent gamma priors are adopted for the parameters, facilitating Bayesian estimation. Due to the non-closed form nature of the mathematical expressions for Bayes and Maximum Likelihood estimators, numerical solutions are provided using the Newton-Raphson and Markov Chain Monte Carlo techniques. Sensitivity analysis is conducted by varying hyperparameters to assess the impact of prior information. Simulation studies, complemented by real-world applications, evaluate the effectiveness of these numerical methods. The results underscore the utility of Bayesian approaches in handling inverse Weibull distribution for survival time upper record values.

Index Term: Record values, Bayes Estimates (BEs), Maximum Likelihood Estimates (MLEs), Bayes Estimates using non-informative prior (BENP), Bayes Estimates using informative prior (BEIP).

MULTIMODAL DATA ANALYSIS

Babar Khan, Tahir Mehmood

School of Natural Sciences (SNS), National University of Sciences and Technology (NUST), Islamabad, Pakistan Author's Email: bravokilo092@gmail.com

In recent years, multimodal data analysis has emerged as a powerful approach to address complex challenges across various domains, including environmental monitoring, industrial safety, and health applications. By integrating data from multiple modalities, such as sensor measurements and thermal images, this approach provides a more comprehensive understanding of the underlying phenomena. However, multimodal data analysis faces significant challenges, such as data noise, feature redundancy, and the complexity of fusion processes. This study explores advanced methodologies to effectively analyze multimodal gas data, focusing on denoising techniques, attention mechanisms, and hybrid learning approaches. Sensor data are denoised using filters like Kalman Filter, Moving Average Filter, and Principal Component Analysis (PCA), while thermal images are enhanced using Non-Local Means (NLM), Gaussian Filtering, and deep learning-based denoising techniques. Attention mechanisms, including self-attention and cross-attention, are employed to prioritize critical features from diverse modalities, ensuring more focused and meaningful data integration. Both classification and non-classification approaches, such as clustering algorithms (K-Means, DBSCAN) and anomaly detection methods (Autoencoders, Isolation Forest) are applied to evaluate their performance in detecting gas anomalies. The findings demonstrate that multimodal data analysis, when enhanced with denoising and attention mechanisms, significantly improves the accuracy and robustness of anomaly detection systems. This work highlights the advantages of integrating multiple data sources, offering practical solutions for safety-critical applications where precision and reliability are paramount.

Keywords: Multimodal Data, Denoising Techniques, Attention Mechanisms, Clustering, Anomaly Detection

Identification of Defects in the Production of Powered Window Regulators using Deep Learning on Vibration and Noise Data

*Asad Riaz, Luigi Carassale

Dynamo Lab, Department of Mechanical, Energy, Management & Transportation Engineering (DIME), University of Genova, Italy

**Corresponding Author's Email: asad.riaz@edu.unige.it*

Powered window regulator is an important component of automobile. Its smooth operation and timely response are crucial for the better functionality of automobile. Defects like noisy operation and slow response imply manufacturing faults. Conventional quality inspection uses auditory assessment method, which is labor insensitive and error prone. This study suggests a deep learning based automated fault detection using a ResNet-50 Convolutional Neural Network (CNN) to classify power window regulator using spectrograms derived from vibration signals. An accelerometer is employed to collect vibration data and labeled as OK or NOK. The time series vibration data is transformed into spectrograms via Short Time Fourier Transform (STFT). The fined tuned ResNet-50 model with a custom classification layer, achieved 96% validation accuracy and 0.066 loss. The results show the model is able to accurately classify noise patterns. It eliminates human error and increases productivity by reducing inspection time 60%. This approach aligns with Industry 4.0 standards, providing a scalable solution for quality assurance in manufacturing. In future work, the model will be deployed on edge devices and will also introduce the root cause analyses. A web application will also be developed for remote supervision, defect sorting, and alerts for recurring issues.

Keywords: Powered Window Regulator, Vibration & Noise, Defects Detection, Deep Learning, ResNet-50, Convolutional Neural Network

Poverty Indicators as Determinants of Chronic Poverty in Punjab: Evidence from Household Data

Muhammad Abdullah Hasni*, Zahoor Ahmad
Department of Statistics, University of Sargodha, Sargodha, Pakistan

*abdullahhasni13@gmail.com

This study explores the effect and contribution of different poverty indicators on multidimensional poverty in Punjab, Pakistan, through the MPI and CMPI. Using panel data from Multiple Indicator Cluster Survey (2011, 2014, 2018), the study assesses the role of major dimensions such as health, education, and living standards in determining the level of poverty. Major regional disparities, with divisions like D.G. Khan and Sargodha perpetually plagued by chronic poverty owing to entrenched deprivation, are identified by the analysis; while other divisions, for instance, Rawalpindi, face transient poverty associated with transitory hardships. Major indicators of multidimensional poverty have been found to have disproportionately large contributions from such indicators as child mortality, educational attainment, and sanitation access. The Alkire-Foster methodology is used in the research to quantify poverty intensity, deprivation levels, and regional variations, thus providing a robust framework for understanding chronic and transient poverty dynamics. The study highlights the importance of region-specific and indicator-specific poverty alleviation strategies. Educational deficits in D.G. Khan, health improvement in Gujranwala, and infrastructure in Sargodha are ranked as the most important priorities. The findings of the study shall guide policymakers in designing data-driven interventions that reduce poverty sustainably and effectively, consistent with national and global development objectives.

Keywords: Multidimensional Poverty Index (MPI), Chronic Poverty, Transient Poverty, Poverty Indicators, Sustainable Development Goals (SDGs), Regional Development

Integrating Latent Variables with Nonlinear Models for Improved High-Dimensional Chemometric Predictions

Sughra Sarwar*, Tahir Mehmood

School of Natural Sciences, Natural University of Sciences and Technology, Islamabad, Pakistan

Email: sugkhan444@gmail.com

The aim of our study is to investigate the integration of latent variables with nonlinear models to enhance chemometric predictions in high-dimensional data settings. Integrating latent variable methods with nonlinear modeling offers a synergistic approach to chemometric analysis. By combining Partial Least Squares (PLS) regression and Multivariate Adaptive Regression Splines (MARS), the approach aims to improve model interpretability and robustness. The integrated model demonstrates improved predictive performance and practical implications in chemometric analysis.

Keywords: High dimensional data, Partial least squares regression, Root mean square error, Mean Absolute error

QSAR Analysis of Certain Degree-Based Topological Descriptors ANN

Aiman

University of Agriculture, Faisalabad.

Email: emantehreem574@gmail.com

This study examines the application of degree-based topological descriptors in QSAR modeling using Artificial Neural Networks (ANNs). Descriptors such as Zagreb and Randic indices are evaluated to predict molecular biological activity. The ANN approach effectively captures non-linear relationships, enhancing prediction accuracy. Results emphasize the significance of these descriptors in QSAR, providing a computationally efficient method for molecular activity prediction.

Measuring the Performance of Supervised Machine Learning Algorithms for Optimizing Productivity Prediction

Azhar Ali

University of Agriculture, Faisalabad.

Email: abcdf4534@gmail.com

This study investigates the performance of supervised machine learning algorithms for productivity prediction, a critical factor in enhancing efficiency across industries. Popular algorithms such as linear regression, decision trees, support vector machines (SVM), and ensemble methods like random forests are evaluated on a dataset containing productivity-related features. Model performance is analyzed using metrics including mean absolute error (MAE), root mean square error (RMSE), and R^2 score, highlighting their predictive accuracy and reliability. The findings reveal the strengths and weaknesses of each algorithm in managing complex, multidimensional data. These insights provide a foundation for selecting the most suitable machine learning model to optimize productivity forecasting and inform data-driven decision-making.

AI-Driven Predictive Modeling for Pancreatic Cancer Detection and Treatment

Muhammad Bilal Khan

University of Agriculture, Faisalabad.

Email: drmbilal172@gmail.com

Pancreatic cancer, one of the deadliest malignancies, demands transformative AI solutions for early detection and precision treatment. This study employs advanced methodologies—Transformers, GANs, and anomaly detection—to integrate multimodal data, including EHRs, CT scans, and biomarkers, achieving unmatched diagnostic precision. Predictive models with an AUC-ROC of 0.88 identify high-risk patients, while GAN-based synthetic CT generation revolutionizes radiotherapy planning by enhancing accuracy and reducing delays. By embedding scalability into machine learning frameworks, this research improves decision-making, minimizes errors, and sets a benchmark for efficient, patient-centered cancer care.

Solar Power Generation: Data Insights and Trends

Mian Fahad Hussain, Tahir Mehmood

School of Natural Sciences (SNS), National University of Sciences and Technology (NUST), Islamabad, Pakistan
Email: fahadasad822@gmail.com

This research focuses on developing advanced predictive models to enhance solar power generation forecasting and evaluate its influence on grid stability. By employing sophisticated methodologies like time series analysis and machine learning, the study aims to address challenges in renewable energy integration. Analytical approaches such as ARIMA, Fourier transforms, and hybrid machine learning models like Random Forest and Gradient Boosting Machines (GBM) are explored. The study further includes classification techniques like Support Vector Machines (SVM) and Decision Trees to optimize predictive accuracy and interpret energy flow dynamics. The outcomes include precise forecasting models for solar energy generation, improved grid stability management, and actionable insights for mitigating voltage fluctuations. These advancements aim to foster sustainable energy solutions and enhance the efficiency of renewable energy grids.

Keywords: Solar energy forecasting, Time series analysis, Machine learning, Grid stability, Renewable energy integration, Voltage fluctuation mitigation

Machine Learning, Applications and its Types

Esha Shahzad

University of Agriculture, Faisalabad.

Email: eshashahzad916@gmail.com

Machine Learning (ML) is a transformative branch of artificial intelligence that enables systems to learn and improve from experience without explicit programming. This study provides an overview of ML types, including supervised, unsupervised, and reinforcement learning, along with their distinctive characteristics. Applications span diverse domains such as healthcare, finance, and automation, showcasing their ability to solve complex problems. By leveraging ML, industries achieve innovation, efficiency, and predictive accuracy, underscoring its pivotal role in shaping the future of technology.

Comparison of SVD and Principal Component Analysis (PCA) based on Image Processing

Muskan Nisar

University of Agriculture, Faisalabad.

Email: muskannisar48@gmail.com

This study compares Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) in the context of image processing. Both techniques reduce dimensionality and improve computational efficiency, but they differ in mathematical approaches and applications. The comparison highlights their effectiveness in image compression, noise reduction, and feature extraction. Results reveal the advantages of PCA in preserving variance and SVD's superior adaptability in low-rank approximations. The findings provide insights for selecting the appropriate method based on specific image processing requirements.

Image Processing using Principal Component Analysis (PCA)

Minahil Zia

University of Agriculture, Faisalabad.

Email: minahilzia555@gmail.com

Principal Component Analysis (PCA) is a powerful dimensionality reduction technique widely used in image processing. This study explores PCA's application in compressing and enhancing images by transforming high-dimensional data into lower-dimensional spaces while preserving critical features. PCA efficiently reduces noise, improves image quality, and accelerates computational processes in tasks like object recognition and facial analysis. The results demonstrate PCA's effectiveness in optimizing image data storage and processing, making it indispensable in modern image analysis workflows.

Air Pollution Forecasting through RNN

Hafiza Raeesa Sohail

University of Agriculture, Faisalabad.

Email: raeesasohailch@gmail.com

Air pollution forecasting is crucial for mitigating environmental and health impacts. This study employs Recurrent Neural Networks (RNNs) to predict air quality by analyzing temporal patterns in pollutant concentration data. RNNs excel at handling sequential data, capturing complex dependencies over time. The model demonstrates high accuracy in forecasting key pollutants like PM2.5, enabling proactive measures to address pollution. This approach highlights the potential of deep learning in environmental monitoring and sustainable urban planning.

Integrating Frechet Distribution with Machine Learning Models for Enhanced Prediction of Extreme Events: A Bayesian Approach

Asim Ali, Zahoor Ahmad

University of Sargodha, Pakistan

Email: zahoor.ahmad@uos.edu.pk; asim316315uos.edu.pk@gmail.com

This article explores the integration of the Fréchet distribution with Bayesian inference and machine learning models to predict extreme events in real-world datasets. The Fréchet distribution, a type of extreme value distribution, is widely used to model the largest values in datasets. In this study, we employ Bayesian inference using a Gamma distribution as a prior to update the scale parameter of the Fréchet distribution based on observed data. We also combine machine learning algorithms, such as XGBoost, Extra Trees, Gradient Boosting, SVR, and RNN (LSTM), to model and predict extreme values. The performance of these models is evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to assess their effectiveness in forecasting extreme events. The results highlight the strengths and weaknesses of each model, showing how machine learning techniques can enhance the prediction of extreme values, compared to traditional methods.

Keywords: Fréchet distribution, Bayesian inference, extreme events, machine learning, prediction.