PROCEEDINGS

The 20th International and National Conference on Applied Computer Technology and Information Systems (ACTIS) and The International and National Conference on Business Administration (NCOBA) 2023-2

Journal & Proceeding ACTIS Applied Computer Technology and Information System

AiAT ARTIFICIAL INTELLIGENCE ASSOCIATION OF THAILAND

Online conference
25 August 2023

http://conference.rpu.ac.th/actis2023

สารจากคณบดีคณะเทคโนโลยีสารสนเทศและนวัตกรรมดิจิทัล มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

งานประชุมวิชาการระดับชาติและระดับนานาชาติด้านเทคโนโลยีคอมพิวเตอร์ประยุกต์และระบบสารสนเทศ ครั้งที่ 20 (ACTIS2023: The 20th International conference in Applied Computer Technology and Information System Acronym) และงานประชุมวิชาการระดับชาติและระดับนานาชาติ ด้านบริหารธุรกิจ ครั้งที่ 20 (NCOBA2023: The 20th National Conference on Business Administration) จัดขึ้นในวันที่ 25 สิงหาคม พ.ศ. 2566 โดยคณะเทคโนโลยีสารสนเทศและนวัตกรรมดิจิทัล มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ คณะเทคโนโลยีดิจิทัล มหาวิทยาลัยราชพฤกษ์ สมาคมปัญญาประดิษฐ์ประเทศไทย และเครือข่ายความร่วมมือในการจัดงานประชุมวิชาการ จำนวน 11 สถาบันการศึกษา งานประชุมวิชาการนี้เป็นอีกหนึ่งเวทีในการแลกเปลี่ยนแนวความคิด องค์ความรู้ สร้างนวัตกรรมด้านเทคโนโลยีคอมพิวเตอร์ ระบบสารสนเทศประยุกต์ เทคโนโลยีดิจิทัล และบริหารธุรกิจ มีวัตถุประสงค์เพื่อร่วมสร้างความเข้มแข็งทางวิชาการในการจัดบริการวิชาการ การทำวิจัย และการพัฒนาบุคลากร เพื่อการพัฒนาประเทศชาติอย่างยั่งยืน

งานประชุมวิชาการครั้งนี้ต้องขอขอบคุณทุกท่านที่ได้ส่งบทความเพื่อเข้าสู่กระบวนการการพิจารณาซึ่งเป็นบทความที่มีคุณภาพสูงทำให้งานประชุมวิชาการได้รักษาคุณภาพและมาตรฐานของงาน ขอขอบพระคุณคณะกรรมการพิจารณาบทความทุกท่านที่เสียสละเวลาอันมีค่าเพื่อให้ได้ผลงานวิจัยที่มีคุณภาพได้นำเสนอในงานประชุมวิชาการในครั้งนี้ด้วย

ขอขอบคุณเครือข่ายความร่วมมือทางเทคโนโลยีคอมพิวเตอร์ประยุกต์และระบบสารสนเทศในประเทศทั้ง 11 สถาบันการศึกษา ที่ให้การสนับสนุนเป็นอย่างดีทั้งด้านคณาจารย์ บุคลากรในการร่วมเป็นคณะกรรมการเครือข่าย คณะกรรมการดำเนินงาน และคณะกรรมการพิจารณาบทความ ได้แก่ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี มหาวิทยาลัยเทคโนโลยีราชมงคลกรุงเทพ มหาวิทยาลัยราชภัฏจันทรเกษม มหาวิทยาลัยสุโขทัยธรรมาธิราช มหาวิทยาลัยทักษิณ มหาวิทยาลัยราชพฤกษ์ มหาวิทยาลัยกรุงเทพสุวรรณภูมิ วิทยาลัยเซาธ์อีสท์บางกอก และมหาวิทยาลัยราชภัฏนครปฐม

ท้ายนี้หวังเป็นอย่างยิ่งว่างานประชุมวิชาการ ACTIS และ NCOBA จะเป็นเวทีในการนำเสนอผลงานวิจัยและนำไปสู่การพัฒนาตนเองเพื่อให้เป็นนักวิจัยที่ดีและมีคุณภาพสูงยิ่ง ๆ ขึ้นไป ACTIS และ NCOBA จะเป็นแพลตฟอร์มสำหรับการแลกเปลี่ยนความรู้ในด้านเทคโนโลยีคอมพิวเตอร์ ระบบสารสนเทศประยุกต์ เทคโนโลยีดิจิทัล และบริหารธุรกิจ สร้างแรงบันดาลใจให้กับนักวิจัยเพื่อคิดค้นนวัตกรรมและพัฒนาองค์ความรู้ใหม่ ๆ เพื่อการพัฒนาชาติอย่างยั่งยืนตลอดไป

(ผู้ช่วยศาสตราจารย์ ดร.สุนันฑา สดสี)
คณบดีคณะเทคโนโลยีสารสนเทศและนวัตกรรมดิจิทัล
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

Message from Dean, Faculty of Information Technology and Digital Innovation, KMUTNB

The 20th International Conference in Applied Computer Technology and Information System Acronym (ACTIS2023) and the 20th National Conference on Business Administration (NCOBA2023), held on 25 August, 2023. ACTIS & NCOBA are organized by Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Faculty of Digital Technology, Rajapruk University, Artificial Intelligence Association of Thailand, and eleven academic collaborations.

The conferences provide a unique platform both for exchanging innovative research in Applied Computer Technology, Information Systems, and Business Administration which aim to enhance collaboration network, researches, and improve the sustainability of our nation.These conferences made huge possible with innovative and progressive contributions from the research community. Herein, special thanks are due to all technical committee members for their diligent consideration of all submissions and for maintaining and preserving the high standards for which ACTIS & NCOBA are justifiably renowned. Regrettably, as a consequence of our rigorous peer review process, we were unable to accept some articles for presentation.

As well as, a debt of gratitude is owed to the eleven co-operative universities for their support and the contributions of staff members by serving in the network committees, executive director committees, and technical program committees. These universities are King Mongkut's University of Technology North Bangkok (KMUTNB), Rajamangala University of Technology Suvarnabhumi (RUS), Rajamangala University of Technology Thanyaburi (RMUTT), Rajamangala University of Technology Krungthep (RMUTK), Chandrakasem Rajabhat University (CRU), Sukhothai

Sunantha Sodsee

Assistant Professor Dr. Dr.-Ing. Sunantha Sodsee
Dean, Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok

สารจากคณบดีคณะเทคโนโลยีดิจิทัล มหาวิทยาลัยราชพฤกษ์

การประชุมวิชาการในครั้งนี้ มีวัตถุประสงค์เพื่อเป็นเวทีระดับชาติ และนานาชาติ ให้นักวิชาการ นักวิจัย คณาจารย์ นิสิต และนักศึกษา ได้เผยแพร่ผลงาน ได้รับฟังและแลกเปลี่ยนองค์ความรู้ ผ่านบทความวิจัยและนวัตกรรม ซึ่งเป็นประโยชน์ ต่อการพัฒนาประเทศ เพื่อสร้างเครือข่ายความร่วมมือพัฒนาความก้าวหน้าทางวิชาการและวิจัยกับสถาบันการศึกษาต่าง ๆ

จากความตระหนักในภารกิจของสถาบันอุดมศึกษาที่ต้องวิจัยเพื่อสร้าง องค์ความรู้ใหม่ด้านเทคโนโลยีคอมพิวเตอร์และระบบสารสนเทศประยุกต์รวมถึงการบริหารธุรกิจอันจะส่งผลต่อการพัฒนาความเจริญก้าวหน้าของประเทศ คณะเทคโนโลยี สารสนเทศและนวัตกรรมดิจิทัล มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ และคณะเทคโนโลยีดิจิทัล มหาวิทยาลัยราชพฤกษ์ ได้รับมอบหมายจากเครือข่ายเป็นเจ้าภาพร่วมจัดการประชุมวิชาการระดับชาติและนานาชาติ ด้านเทคโนโลยีคอมพิวเตอร์และระบบสารสนเทศประยุกต์ (ACTIS) การประชุมวิชาการระดับชาติและนานาชาติ ด้านบริหารธุรกิจ (NCOBA) ครั้งที่ 20 ซึ่งการประชุมวิชาการในครั้งนี้ได้รับความร่วมมือจากสมาคมปัญญาประดิษฐ์ ประเทศไทย

สุดท้ายนี้ ขอขอบคุณเครือข่ายการประชุมวิชาการ ACTIS และ NCOBA ทั้ง 11 มหาวิทยาลัย และคณะกรรมการจัดงานประชุมวิชาการจากคณะเทคโนโลยีสารสนเทศและนวัตกรรมดิจิทัล มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ และคณะเทคโนโลยีดิจิทัล มหาวิทยาลัยราชพฤกษ์ ที่มีความมุ่งมั่นและความพยายาม เพื่อให้การประชุมวิชาการ ACTIS และ NCOBA ครั้งที่ 20 ปี 2023-2 นี้เป็นงานที่สร้างความประทับใจและน่าจดจำสำหรับผู้เข้าร่วมทุกท่าน

ดิฉันหวังไว้เป็นอย่างยิ่งว่าทุกท่านจะมีช่วงเวลาที่ดี และได้รับประโยชน์จากการเข้าร่วมการประชุมวิชาการ ACTIS และ NCOBA ครั้งที่ 20 ปี 2023-2 และหวังว่าทุกท่านจะเข้าร่วมการประชุมวิชาการ ACTIS และ NCOBA ครั้งที่ 21 ปี 2024 ต่อไป

ผู้ช่วยศาสตราจารย์ ดร.เรวดี ศักดิ์ดุลยธรรม
คณบดีคณะเทคโนโลยีดิจิทัล
มหาวิทยาลัยราชพฤกษ์

Message from the Dean, Faculty of Digital Technology, Rajapruk University

This conference is held with the purposes in being a stage for scholars, researchers, professors, students to present and disseminate their work for public engagement and knowledge exchange through research articles and innovation which will be beneficial for country development and to build a cooperation network on the development of academic achievement and research with other educational institutions.

Due to the awareness of a mission of higher education institutions in conducting a research study to build a new knowledge in the field of applied computer technology and information system including business administration, which shall affect the development and progress of the country, Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok and Digital Technology Faculty, Rajapruk University were assigned from the Network to be the co-host for holding the 20th International and National Conference on Applied Computer Technology and Information Systems (ACTIS) and the International and National Conference on Business Administration (NCOBA). This academic conference received good cooperation from Artificial Intelligence Association of Thailand (AIAT).

Finally, I thank the network of the International and National Conference on ACTIS and NCOBA from all 11 universities and the conference joint-organizing committees from Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok and Digital Technology Faculty, Rajapruk University who have shown great commitment and fantastic team efforts in ensuring that the 20th ACTIS and NCOBA 2023-2 will be an impressive and memorable event for all participants.

I wish all of you a delightful and fruitful time in the 20th ACTIS and NCOBA 2023-2 and I look forward to your participation in the 21th ACTIS and NCOBA 2024.

Raywadee

Assistant Professor Dr. Raywadee Sakdulyatham
Dean, Faculty of Digital Technology
Rajapruk University

คณะกรรมการอำนวยการ
(Steering Committee)

ผู้ทรงคุณวุฒิพิจารณาบทความ
(Paper Reviewer)

Related topics of Conference

ACTIS – Conference Tracks:

Information Technology
1.1 Artificial Intelligence and Machine Learning
1.2 Internet Technologies and Applications
1.3 Data/Network Security
1.4 Storage Systems and Techniques
1.5 Data Retrieval & Data Mining 1.6 System Modeling and Simulations

Commerce Technology
2.1 e-Business Applications and Software
2.2 Managing Innovation & Marketing on the Web
2.3 Web Advertising and Web Publishing
2.4 Business & Consumer Oriented E-Commerce
2.5 Business Technology

Software Engineering
3.1 Software Process, Design and Architecture
3.2 Software Configuration Management
3.3 Software Evolution
3.4 Software Component and Prototyping

Computer Systems
4.1 Computer Systems and Applications
4.2 Distributed & Grid Computing
4.3 Computer Graphics & HCI
4.4 Computer Networks; Protocols & QOS
4.5 Network Management

Computer Education & Training
Computer Hardware and Peripheral
Digital Media & Technology

NCOBA – Conference Tracks:

International Business Administration
1.1 Principle of International Business
1.2 International Business Management
1.3 Business Development
1.4 Innovation and Technology Management

Marketing Management
2.1 Principle of Business Management
2.2 Marketing Environment Analysis
2.3 Marketing Strategy

Finance
3.1 Banking and Finance
3.2 Corporate Finance and Governance
3.3 International Finance

Economics
4.1 Economic Development
4.2 Economic Methodology
4.3 Labour Economics
4.4 International Economics

Table of Articles

The International Conference on Applied Computer Technology and Information Systems (ACTIS)

A Comparative Study between Generalized Linear Models and Generalized Additive Models in the Modeling of Health Biological Signal Data

Natakon Nawaratana[1]
School of Mathematics
Institute of Science
Suranaree University of Technology
Nakhon Ratchasima, Thailand
Natakonnawaratana@gmail.com

Amornrat Suriyawichitseranee[2]
School of Mathematics
Institute of Science
Suranaree University of Technology
Nakhon Ratchasima, Thailand
amornrat@g.sut.ac.th

Jessada Tanthanuch[3]
School of Mathematics
Institute of Science
Suranaree University of Technology
Nakhon Ratchasima, Thailand
jessada@g.sut.ac.th

***Abstract*— The utilization of the Generalized Linear Model (GLM) and Generalized Additive Model (GAM) plays a crucial role in applications of artificial intelligence (AI) and mathematical modeling. However, the GAM surpasses the GLM in terms of its nonlinear generalizability. This research aims to compare the study between GLM and GAM in the modeling of health biological signal data. The dataset used in this study encompasses information about the presence or absence of smoking, obtained from bio-signals. The dataset is sourced from the National Health Insurance Service_Health Checkup Information (Korea) and can be accessed at https://www.data.go.kr/data/15007122/fileData.do. It consists of 22 variables and includes a total of 55,692 records. In the research procedure, the first step involved assessing the correlation among variables in order to reduce the number of variables utilized in the model. Subsequently, the models were constructed considering four distributions: normal, Tweedie, gamma, and inverse-Gaussian distributions. The performance of the models was evaluated based on the Akaike information criterion (AIC), the root mean square error (RMSE) and the distance between indices of simulation and observation (DISO) metrics. The research findings indicate that GAM outperforms GLM overall, as evidenced by lower AIC, RMSE and DISO. The best performing forecasting models for cholesterol and triglyceride levels are the models created by GAM that take into account the normal distribution of the data.**

Keywords*— *generalized linear model, generalized additive model, biological signal data, DISO.

I. INTRODUCTION

The world has experienced severe devastation due to the COVID-19 pandemic, which has resulted in widespread illness, loss of life, and significant economic disruption. Excess weight can elevate the risk of developing severe symptoms and complications associated with the disease, such as pneumonia, blood infections, and cardiovascular issues. Additionally, obesity can negatively affect the body's immune system, potentially making it more difficult to fight off infections. In these challenging times, it is more important than ever to take care of our individual health. Patients who are overweight may have an increased risk of experiencing severe impacts when infected with COVID-19 [1]. Regular health checkups are an important part of maintaining good health. They can help to detect and treat health problems early on, which can help to protect us from the serious consequences of COVID-19 and other health problems. Biological signal data refers to data collected from the human body, which serves to monitor health and detect potential health issues. It encompasses various types of data, including Electrocardiogram (ECG), Electroencephalogram (EEG), Electromyogram (EMG), respiratory monitoring, blood pressure monitoring, as well as measurements of cholesterol and triglyceride levels. These datasets enable healthcare professionals to assess individuals' well-being and identify any underlying medical conditions [2]. Cholesterol and triglycerides are essential components of biological signal data, influencing various physiological processes. Imbalances in their levels can significantly impact overall health and increase the risk of chronic conditions [3]. Mathematics and statistics are crucial in modeling these biomarkers, employing differential equations to predict their behavior over time and statistical models to identify relationships with other variables.

Statistical distributions are essential in statistics for modeling data across various domains, serving a multitude of purposes. They find applications in predicting future outcomes, making informed decisions, assessing event probabilities, and comparing populations. In the realm of biomedical modeling, several widely employed statistical distributions facilitate the analysis and interpretation of data. These distributions, such as the normal distribution, gamma distribution, inverse-Gaussian distribution, and Tweedie distribution, are powerful tools empowering researchers to extract valuable insights, make well-founded decisions, and draw meaningful conclusions from available data [4,5].

Generalized linear models (GLMs) and generalized additive models (GAMs) are powerful statistical tools that go beyond traditional linear models, enabling the modeling of diverse data types, even those that do not follow a normal distribution. GLMs and GAMs offer advanced techniques that enhance modeling accuracy by accommodating both normal and non-normal distributional assumptions and capturing potential linear and nonlinear relationships. While GLMs allow for linear relationships between variables, GAMs excel in capturing non-linear relationships, making them highly versatile for various applications in artificial intelligence (AI). These models find utility in predictive modeling, feature selection, and exploratory data analysis. In the realm of natural language processing (NLP), both GLMs and GAMs are valuable for modeling relationships between words and phrases in text. They facilitate language understanding, sentiment analysis, and text generation tasks. Additionally, in the field of computer vision, these models are essential for understanding intricate relationships between pixels in images. They enhance image classification, object detection, and scene understanding with greater accuracy and depth [6]. A key advantage of GLMs and GAMs is their ability to handle categorical and count data commonly encountered in biological signal analysis within the health field. Their flexibility makes them well-suited for analyzing health data. These models provide valuable insights into understanding and predicting the dynamics of cholesterol and triglyceride levels, benefiting research and clinical applications in the health field.

This research focuses on a comparative study of utilizing GLMs and GAMs to model cholesterol and triglyceride levels using additional biological signal data. Both GLMs and GAMs encompass the utilization of four exponential family distributions: the normal distribution, gamma distribution, inverse-Gaussian distribution and Tweedie distribution. The performance of these models is evaluated using metrics such as Akaike information criterion (AIC), the root mean square error (RMSE), and the distance between indices of simulation and observation (DISO).

II. Related Statistical Distributions

In this research, we focus on 4 four exponential family distributions, which are the normal distribution, gamma distribution, inverse- Gaussian distribution and Tweedie distribution.

A. *Normal Distribution*

The normal distribution, commonly referred to as the Gaussian distribution, is a bell-shaped curve with symmetry that serves as a model for data distribution. In a normal distribution, the mean, median, and mode are all equivalent, and the area beneath the curve sums up to 1. The standard deviation of the normal distribution quantifies the dispersion or spread of the data.

The formula that describes the probability density function (PDF) of a normal distribution is as follows:

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where x is a random variable, μ is the mean of the distribution, and σ is the standard deviation of the distribution [6].

The normal distribution is a powerful tool for modeling, analyzing, and medical statistics [7].

B. *Gamma Distribution*

The gamma distribution, commonly employed for modeling positively skewed continuous data, is characterized by two parameters: the shape parameter (α) and the scale parameter (β). The PDF of the gamma distribution is expressed as follows:

$$f(x;\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$$

where x is a non-negative real number, α is the shape parameter, β is the scale parameter, and $\Gamma(\alpha)$ is the gamma function.

the gamma distribution can be utilized to predict medical outcomes and inform decision-making in healthcare [8].

C. *Inverse-Gaussian Distribution*

The inverse-Gaussian distribution is a continuous probability distribution that was initially introduced by the British statistician Harold Jeffreys in 1935. It gets its name from being the reciprocal of the normal distribution [5].

The PDF of the inverse-Gaussian distribution is given by:

$$f(x;\mu,\lambda)=\frac{a}{\lambda\beta\left(a/2,b/2\right)x}\exp\left(-\frac{b(x-\mu)^2}{\lambda}\right),$$

where, x is a random variable following the inverse-Gaussian distribution, μ is the mean of the distribution, λ is the shape parameter, a and b are constants that depend on μ and λ, and $\beta(\cdot,\cdot)$ is the beta function. The mean and variance of the distribution are μ and μ^2/λ, respectively.

The inverse-Gaussian distribution has found applications in medical research, such as its use in analyzing the factors that impact the survival of patients with oesophageal cancer through parametric analysis with frailty models [9].

D. *Tweedie Distribution*

The Tweedie distribution is a versatile family of probability distributions frequently employed in actuarial, financial, and ecological applications. It is specifically designed to model non-negative data with skewness and heavy tails. Introduced by Maurice Tweedie in 1984, the distribution has evolved to encompass various distributions. The PDF of the Tweedie distribution is expressed as follows:

$$f(x;\mu,\sigma^2,p)=\frac{c(x,\sigma^2,p)}{\Gamma(p)}\exp\left(-\frac{\mu x}{\sigma^p}\right)\left(\frac{y}{\sigma^p}\right)^{p-1},$$

where μ is the mean of the distribution, σ is the dispersion parameter, p is the power parameter, $c(x,\sigma^2,p)$ is a normalizing constant, and $\Gamma(p)$ is the gamma function.

It is intriguing to note that numerous distributions within the Tweedie family are characterized by the range of values for the index parameter. For instance, notable examples include the normal distribution (p=0), the gamma distribution (p=2), and the inverse-Gaussian distribution (p=3) [10].

III. Models

A brief information of GLM and GAM is provided in this section.

A. *Generalized Linear Model*

A generalized linear model (GLM) is a flexible statistical framework that extends linear regression to accommodate various types of response variables and their associated probability distributions. It provides a way to model the relationship between a set of predictors and a response variable while accounting for non-normal and non-continuous data. GLMs differ from traditional linear models in that they do not assume the normal distribution of the response variable and do not strictly require an expected value of the response variable to be a linear combination of explanatory variables. They allow for the use of different link functions to relate the predictors to the response, making them suitable for analyzing a wide range of data types, including binary, count, and categorical outcomes. The flexibility of GLMs makes them widely used in fields such as medicine, social sciences, and economics, where complex data structures and diverse response variables are encountered.

The GLM incorporates a smooth and invertible link function denoted as $g(\cdot)$. This function transforms the expected value of the response variable Y , $\mu = E(Y)$ to the summation:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n,$$

where $E(Y)$ is the expected value of the response variable Y , $X_1, \ldots, X_n$ are explanatory variables, and β_i is the model parameter, $i = 0, \ldots, n$. The proposed equation is a linear predictor, which incorporates information about the independent (explanatory) variables into the model [11].

B. Generalized Additive Model

A generalized additive model (GAM) is a statistical modeling technique that extends the concept of generalized linear models (GLMs) by allowing for nonlinear relationships between predictors and the response variable. Unlike GLMs, which assume linear relationships, GAMs can capture complex and nonlinear patterns by incorporating smooth functions of the predictors. In a GAM, the response variable is still related to the predictors through a specified probability distribution and a link function, similar to GLMs. However, instead of assuming a linear relationship, GAMs employ smoothing functions, such as splines, regression or wavelets, to model nonlinearities and interactions. Compared to the linear predictor of GLM, the GAM model can be described as follows:

$$g(\mu) = \beta_0 + f_1(X_1) + \cdots + f_n(X_n),$$

where $g, \mu, \beta_0, X_1, \ldots, X_n$ are the values defined in the section of Generalized Linear Model, and $f_1, \ldots, f_n$ are smooth functions [12].

By allowing for flexible and nonparametric modeling, GAMs can handle complex data structures and capture intricate relationships between predictors and the response. They are particularly useful when dealing with data that exhibit nonlinear patterns, such as time series, spatial data, and interactions between variables [13].

IV. EVALUATION TOOLS

The three evaluation tools used in this research are presented as follows.

A. Akaike Information Criterion

In statistics, the Akaike Information Criterion (AIC) is a criterion used to select the most appropriate model from a set of models, all fitted to the same data but having different explanatory parameters. The AIC is a measure of the trade-off between model fit and model complexity. It aims to balance the goodness of fit of the model with the number of parameters used in the model. The AIC score is calculated based on the model's likelihood function and the number of parameters used in the model. It is given by the formula

$$\mathrm{AIC} = -2\ln \hat{L} + 2p,$$

where $\hat{L}$ is the likelihood function of model, and p is the number of parameters in the model. However, in the context of parameter estimation using the method of least squares, the AIC can be used to compare different models with different numbers of parameters. The AIC for this model can be calculated using the following formula:

$$\mathrm{AIC} = N \ln\left(\frac{RSS}{N}\right) + 2p,$$

where N is the number of data points (sample size), p is the number of parameters in the model, and RSS is Residual Sum of Squares which measures the sum of the squared differences between the observed data and the predicted values from the model [14].

A lower AIC score indicates a better model, as it suggests that the model fits the data well while using fewer parameters, thus avoiding overfitting. Overfitting occurs when a model is too complex and captures noise in the data rather than the underlying pattern, leading to poor performance on new, unseen data. AIC is commonly used in the analysis of regression models and time series models, where different combinations of explanatory variables or lagged terms are considered. By comparing the AIC scores of various models, researchers can identify the model that strikes the best balance between accuracy and simplicity, ultimately aiding in making informed decisions during model selection.

B. Root Mean Square Error

RMSE, which stands for Root Mean Squared Error, serves as a metric to evaluate prediction accuracy. It is obtained by taking the square root of the average squared error. The mean squared error is computed by summing up the squared differences between the predicted values and the actual values, and then dividing the sum by the number of observations. The RMSE formula is

$$\mathrm{RMSE}(x, y) = \sqrt{\frac{\sum_{i=1}^{N} (x_i - y_i)^2}{N}},$$

where x is the observed value, y is the predicted value, and N is the number of observations. A smaller RMSE value signifies a more accurate prediction [12].

C. DISO

DISO, or Distance Between Indices of Simulation and Observation, serves as a statistical measure that allows for the quantification of the disparity between simulated and observed data. To calculate DISO, both sets of data are normalized to have a mean of zero and a standard deviation of one, and then the Euclidean distance between the two sets is computed. DISO finds its application in comparing different simulation models and assessing the impact of parameter modifications on a model. It also helps in identifying areas where the model's performance is subpar, thus providing valuable insights for further research and development endeavors. A low DISO value signifies a close resemblance between the model's output and the observed data, while a high DISO value indicates a lack of accuracy in capturing the underlying process. DISO is defined as follows:

$$\mathrm{DISO} = \sqrt{(1-R)^2 + \mathrm{NAE}^2 + \mathrm{NRMSE}^2},$$

where R is correlation coefficient, NAE and NRMSE are normalized absolute error (AE) and RMSE, respectively. Note that the formulae of R, NAE and NRMSE are as follows:

$$R = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}}, \mathrm{NAE} = \frac{\sum_{i=1}^{N}|x_i-y_i|}{N\bar{x}},$$

and $\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\bar{x}},$

where x_i represents the observed value, y_i represents the predicted value, $\bar{x}$ and $\bar{y}$ represent the means of x_i and y_i, and N is the total number of data points [15].

V. METHODOLOGY

A. Tools

RStudio was selected as the main software for modeling and statistical analysis because of its extensive capabilities and wide acceptance. The research used RStudio version 3.6.1 on a Microsoft Windows 11 Home Single Language, version 22H2 operating system. The computations were executed on a computer equipped with an Intel(R) I7-12700H CPU.

B. Data Set

The dataset is sourced from the National Health Insurance Service_Health Checkup Information provided by National Health Insurance Corporation (Korea) and can be accessed at https://www.data.go.kr/data/15007122/fileData.do. It was first registered on 29 September 2021 and was first published on 19 December 2022. Health check-up information pertains to the overall health examination outcomes of Korean national health insurance employees, dependents aged 40 and above, local subscribers who are household heads, and local subscribers aged 40 and above. It also encompasses individuals who have attained the ages of 40 and 66, among those who are eligible for general health check-ups. The data consists of 22 variables as shown in table 1. It includes a total of 55,692 records that are updated yearly.

C. Model Creation

In this research, we consider the GLMs and GAMs for Y_1 (cholesterol) and Y_2 (triglyceride). First, we assume that Y_1 depends on Y_2, x_1, x_2, ..., x_{20} and Y_2 depends on Y_1, x_1, x_2, ..., x_{20}. Subsequently, the model creations adhere to the provided algorithm.

Table 1. Variables in the data set.

Variable	Description
Y_1	cholesterol (total)
Y_2	triglyceride
X_1	age (years, 5-years gap)
X_2	height (cm)
X_3	weight (kg)
X_4	waist (cm)
X_5	eyesight (left)
X_6	eyesight (right)
X_7	hearing (left)
X_8	hearing (right)
X_9	systolic (blood pressure)
X_{10}	relaxation (blood pressure)
X_{11}	fasting blood sugar
X_{12}	HDL (High-density Lipoprotein)
X_{13}	LDL (Low-density Lipoprotein)
X_{14}	hemoglobin
X_{15}	urine protein
X_{16}	serum creatinine
X_{17}	AST (Aspartate transaminase)
X_{18}	ALT (Alanine transaminase)
X_{19}	GGTP
X_{20}	dental caries

Model Creation Algorithm

1. Importing Data.
2. Selecting a statistical distribution for model creation.
3. Creating the model (using GLM or GAM).
4. Performing feature selection using variable selection methods.
5. Applying the model obtained in step 4. to the test data for the prediction.
6. Evaluating performance.

VI. Research Results

A. Evaluation of the Models for Predicting Cholesterol

Table 2. Results of the performance evaluation of the GLM models for predicting cholesterol.

Distribution	AIC	RMSE	DISO
Normal	347705.3	**16.00829**	**0.1070061**
gamma	275727.2	19.96705	0.4578677
inverse-Gaussian	319671.9	16.53430	0.5317749
Tweedie	**271367.8**	29.70599	1.6851250

Table 3. Results of the performance evaluation of the GAM models for predicting cholesterol.

Distribution	AIC	RMSE	DISO
Normal	**230529.0**	**3.650386**	**0.01387663**
gamma	306760.2	4.355414	0.01731654
inverse-Gaussian	239933.4	4.682186	0.01977855
Tweedie	242879.0	6.221848	0.02663219

B. Evaluation of the Models for Predicting Triglyceride

Table 4. Results of the performance evaluation of the GLM models for predicting triglyceride.

Distribution	AIC	RMSE	DISO
Normal	419560.9	**49.91192**	**0.7328266**
gamma	419536.1	63.69959	0.9537604
inverse-Gaussian	419758.6	66.26119	0.9768965
Tweedie	**418488.8**	110.3757	1.0656720

Table 5. Results of the performance evaluation of the GAM models for predicting triglyceride.

Distribution	AIC	RMSE	DISO
Normal	**350369.8**	**18.73588**	**0.1844699**
gamma	419536.1	61.50394	0.9224283
inverse-Gaussian	416590.5	63.24360	0.9758391
Tweedie	416518.2	61.23025	0.9209138

Based on Tables 2 to 5, they show that GAMs provide overall better performing models than GLMs. Considering RMSE and DISO, both types of models yield the best results for the normal distribution. The models obtained using GLM and GAM techniques for the normal distribution are presented in Tables 6 to 9.

C. GLMs and GAMs Obtained by Using the Normal Distribution

Table 6. The coefficients of GLM in constructing a predictive model for cholesterol (using the Normal distribution).

Coeff	Est	Std. Err.	*t* val.	Pr(>|*t*|)
Intercept	79.410295	2.929922	27.103	< 2e-16
Y_2	0.190650	0.001767	107.888	< 2e-16
X_1	0.057512	0.010598	5.427	5.77e-08
X_2	-0.290947	0.017919	-16.237	< 2e-16
X_3	0.154368	0.012734	12.123	< 2e-16
X_9	-0.083214	0.012256	-6.790	1.14e-11
X_{10}	0.127845	0.017124	7.466	8.46e-14
X_{11}	-0.051593	0.005469	-9.434	< 2e-16
X_{12}	0.935035	0.008429	110.928	< 2e-16
X_{13}	0.602160	0.002499	240.932	< 2e-16
X_{14}	0.787507	0.085317	9.230	< 2e-16
X_{17}	-0.020446	0.007986	-2.560	0.010461
X_{18}	0.016865	0.004910	3.435	0.000594
X_{19}	-0.007051	0.002477	-2.847	0.004414

Remark: The coefficient parameters of X_4, X_5, X_6, X_7, X_8, X_{15}, X_{16}, and X_{20} in the GLM proposed in Table 6 have been discarded in the feature selection process, indicating that these variables are not significant for the model.

Table 7. The Anova for Parametric Effects of GAM in constructing a predictive model for cholesterol (using the Normal distribution).

$f(\cdot)$ Function of	Sum and Mean Sq.	*F* val.	Pr(>*F*)
Y_2	2,923,557	1.3522e+05	< 2.2e-16
X_1	83,105	3.8439e+03	< 2.2e-16
X_2	126,944	5.8716e+03	< 2.2e-16
X_3	229,621	1.0621e+04	< 2.2e-16
X_4	12,882	5.9582e+02	< 2.2e-16
X_5	1,725	7.9808e+01	< 2.2e-16
X_6	248	1.1483e+01	0.0007031
X_9	74,472	3.4446e+03	< 2.2e-16
X_{10}	148,720	6.8788e+03	< 2.2e-16
X_{11}	1,816	8.3981e+01	< 2.2e-16
X_{12}	5,040,221	2.3313e+05	< 2.2e-16
X_{13}	26,259,130	1.2146e+06	< 2.2e-16
X_{14}	83	3.8488e+00	0.0497897
X_{15}	95	4.4021e+00	0.0359017
X_{16}	13	5.8130e-01	0.4457965
X_{17}	9	3.9370e-01	0.5303631
X_{18}	6	2.9190e-01	0.5890280
X_{19}	10	4.4580e-01	0.5043250

Remark: The parameters X_7, X_8, and X_{20} in the GAM proposed in Table 7 have been discarded during the feature selection process, indicating that they are not significant for the model.

Table 8. The coefficients of GLM in constructing a predictive model for triglyceride (using the Normal distribution).

Coeff	Est	Std. Err.	*t* val.	Pr(>|*t*|)
Intercept	-29.349710	8.209824	-3.575	0.000351
Y_1	1.204597	0.011168	107.860	< 2e-16
X_2	-0.109595	0.047531	-2.306	0.021129
X_3	0.117862	0.051194	2.302	0.021326
X_4	0.458089	0.057983	7.900	2.85e-15
X_5	1.295903	0.541842	2.392	0.016777
X_9	0.118274	0.030543	3.872	0.000108
X_{10}	0.259282	0.043023	6.027	1.69e-09
X_{11}	0.328004	0.013495	24.305	< 2e-16
X_{12}	-2.344657	0.021223	110.480	< 2e-16
X_{13}	-0.743345	0.009171	-81.058	< 2e-16
X_{14}	2.691483	0.217812	12.357	< 2e-16
X_{16}	-3.603069	1.341515	-2.686	0.007238
X_{17}	-0.155830	0.020002	-7.791	6.82e-15
X_{18}	0.051627	0.012307	4.195	2.73e-05
X_{19}	0.251567	0.006110	41.174	< 2e-16

Remark: The coefficient parameters of X_1, X_6, X_7, X_8, X_{15}, and X_{20} in the GLM proposed in Table 8 have been discarded in the feature selection process, indicating that these variables are not significant for the model.

Table 9. The Anova for Parametric Effects of GAM in constructing a predictive model for triglyceride (using the Normal distribution).

$f(\cdot)$ Function of	Sum and Mean Sq.	*F* val.	Pr(>*F*)
Y_1	136,584,895	2.9206e+05	< 2.2e-16
X_1	460,637	9.8499e+02	< 2.2e-16
X_2	4,865,888	1.0405e+04	< 2.2e-16
X_3	27,687,582	5.9205e+04	< 2.2e-16
X_4	2,432,383	5.2012e+03	< 2.2e-16
X_5	40,033	8.5604e+01	< 2.2e-16
X_6	8,796	1.8809e+01	1.448e-05
X_9	723,778	1.5477e+03	< 2.2e-16
X_{10}	2,371,565	5.0712e+03	< 2.2e-16
X_{11}	1,598,055	3.4172e+03	< 2.2e-16
X_{12}	53,279,944	1.1393e+05	< 2.2e-16
X_{13}	202,500,607	4.3301e+05	< 2.2e-16
X_{14}	40,658	8.6940e+01	< 2.2e-16
X_{15}	4,405	9.4204e+00	0.002147
X_{16}	1,632	3.4896e+00	0.061761
X_{17}	16,436	3.5146e+01	3.084e-09
X_{18}	2,059	4.4036e+00	0.035870
X_{19}	382,813	8.1858e+02	< 2.2e-16

Remark: The parameters of X_7, X_8, and X_{20} in the GAM proposed in Table 9 have been discarded during the feature selection process, indicating that they are not significant for the model.

D. Explanation of Acquired GLMs and GAMs

As proposed in Table 6-9, some features of the GLMs and the GAMs, have been discarded during the feature selection process. This suggests that these features are not significant for the models. In Tables 6 and 8, the "Est" (Estimate) signifies the coefficient value of the predictor variables, "Std. Err." (Standard Error) quantifies the variability of the coefficient estimate, "*t* val" (*t*-value) indicates the number of standard errors the coefficient estimate deviates from zero, and "Pr(>|*t*|)" (*p*-value) represents the probability linked with observing a *t*-value. Consequently, the absolute *t*-values of all predictors are noticeably distant from zero, suggesting a possibly pronounced impact of the predictors. Furthermore, *p*-values less than 0.05 affirm the statistical significance of the coefficients. Table 6 reveals that AST holds the lowest level of significance among the features considered in the GLM of the predictive cholesterol model. Similarly, Table 8 demonstrates that height, weight, and eyesight (left) are the three least significant features in GLM of the predictive model for triglyceride levels.

Within the framework of the GAM model structure, the ANOVA technique was harnessed to scrutinize the model. Within Tables 7 and 9, the terms "Sum and Mean Sq" portray the accumulation of squared disparities between observed values and the overall mean, while also reflecting the mean of these squared differences. These statistics are instrumental in evaluating the model's variance components. "*F* val" (*F* value) serves to scrutinize whether the means of diverse groups exhibit significant dissimilarity. Furthermore, "Pr(>*F*)" (*p*-value) signifies the likelihood of encountering the computed *F* value when the assumption of insignificant variance between group means holds true. These parameters collectively gauge the significance of the comprehensive model and the effects of individual factors within it. The observed Pr(>*F*) values, akin to Pr(>|*t*|), validate the statistical significance of these predictors. In Table 7, it is evident that the features serum creatinine, AST, ALT, and GTP do not hold statistical significance, while hemoglobin and urine protein are of lesser significance within the context of the GAM for predicting cholesterol levels. Furthermore, Table 9 highlights that serum creatinine lacks statistical significance, while ALT is of lesser significance in the GAM of the predictive model for triglyceride levels.

VII. Conclusion

By the aforementioned data, it is evident that applying the normal distribution in constructing predictive models for cholesterol and triglyceride using the statistical techniques of GLM and GAM yields more efficient models compared to other statistical distributions, gamma distribution, inverse-Gaussian distribution, and Tweedie distribution. The ability of hearing and dental caries status does not affect the development of models for predicting cholesterol and triglyceride levels. In addition to the mentioned explanatory variables, GAMs utilize all the remaining variables, whereas GLMs use fewer explanatory variables. In general, GAMs perform better in model creation compared to GLMs. Based on the data provided, it is apparent that the relationships between the two response variables (cholesterol and triglyceride) and the explanatory variables are nonlinear. Additionally, the data demonstrates a statistically normal distribution. Although GAMs have the capability to create superior forecasting models, interpreting the results from the smoothing functions of each explanatory variable proves to be challenging. However, it is still possible to observe the significant impact of the explanatory variables on both response variables.

ACKNOWLEDGMENT

Authors wishing to acknowledge School of Mathematics, Institute of Science, Suranaree University of Technology, Thailand. This research and researchers have a financial support by the Development and Promotion of Science and Technology Talents Project (DPST scholarship.

REFERENCES

[1] W. Sawadogo, M. Tsegaye, A. Gizaw, and T. Adera, "Overweight and obesity as risk factors for COVID-19-associated hospitalizations and death: Systematic review and meta-analysis," BMJ Nutrition, Prevention & Health, vol. 0, no. e000375, Jan. 18, 2022. doi: 10.1136/bmjnph-2021-000375.

[2] S. Bernhard, A. Brensing, and K.-H. Witte, "Biosignal Processing: Fundamentals and Recent Applications with MATLAB®," De Gruyter Oldenbourg, 2022.

[3] R. McDermid, "Statistics in medicine," Anaesthesia & Intensive Care Medicine, vol. 22, no. 7, pp. 454-462, 2021. doi: 10.1016/j.mpaic.2021.05.009.

[4] C. Forbes, M. Evans, N. Hastings, and B. Peacock, "Statistical Distributions, Fourth Edition," Wiley, 2011.

[5] K. Krishnamoorthy, "Handbook of Statistical Distributions with Applications," Chapman & Hall/CRC, 2006.

[6] A. J. Dobson, "Applied Generalized Linear Models," Chapman and Hall/CRC, 2002.

[7] A. Agresti, "An introduction to categorical data analysis," 2nd ed., New York, NY, USA: Wiley, 2002.

[8] C. J. Cleophas and A. H. Zwinderman, "Gamma Distribution for Estimating the Predictors of Medical Outcome Scores (110 Patients," in Machine Learning in Medicine - A Complete Overview, Springer, Cham, 2020, doi: 10.1007/978-3-030-33970-8_85.

[9] M. R. Ghadimi, M. Mahmoodi, K. Mohammad, M. Rasouli, H. Zeraati, and A. Fotouhi, "Factors affecting survival of patients with oesophageal cancer: a study using inverse Gaussian frailty models," Singapore Med J, vol. 53, no. 5, pp. 336-343, 2012.

[10] B. Jorgensen, "Exponential Dispersion Models," J. R. Statist. Soc. B, vol. 49, no. 2, pp. 127-146, 1987.

[11] J. K. Lindsey, "Applying Generalized Linear Model," Springer-Verlag, 1997.

[12] S. N. Wood, "Generalized Additive Models: An Introduction With R," 2nd ed., Chapman and Hall/CRC Press, 2017.

[13] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2nd ed., Springer, 2009.

[14] Office of the Civil Service Commission, "Dictionary of Statistical Terms, Royal Institute Edition," 2nd revised ed., Bangkok, Thailand: Office of the Civil Service Commission, 2018, ISBN: 9786163890832.

[15] Q. Zhou, D. Chen, Z. Hu, and X. Chen, "Decompositions of Taylor diagram and DISO performance criteria," International Journal of Climatology, vol. 41, no. 1, pp. 1-7, 2021. doi:10.1002/joc.7149.

20th ACTIS & NCOBA

20th International and National Conference on Applied Computer Technology and Information Systems (ACTIS) and the International and National Conference on Business Administration (NCOBA).

The 20th International and National Conference

on Applied Computer Technology and Information System (ACTIS) and Business Administration (NCOBA)

This certificate is presented to

Natakon Nawaratana, Amornrat Suriyawichitseranee and Jessada Tanthanuch

For The Paper Titled

A Comparative Study between Generalized Linear Models and Generalized Additive Models in the Modeling of Health Biological Signal Data

25 August 2023

Raywadee

Raywadee Sakdulyatham, Ph.D.
Dean of Faculty of Digital Technology

The 20th International and National Conference

on Applied Computer Technology and Information System (ACTIS) and Business Administration (NCOBA)

BEST PAPER AWARD

awarded to

BEST PAPER

Natakon Nawaratana, Amornrat Suriyawichitseranee and Jessada Tanthanuch

For The Paper Titled

A Comparative Study between Generalized Linear Models and Generalized Additive Models in the Modeling of Health Biological Signal Data.

25 August 2023

Raywadee

Raywadee Sakdulyatham, Ph.D.
Dean of Faculty of Digital Technology