Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/140204

Predictive Modeling of Breast Cancer Incidence: A Comparative Study of Fuzzy Time Series and Machine Learning Techniques

Ibtisam Daqqa[1], *Abdullah M. Almarashi*[2], *Mnahil. M. Bashier*[3], *M. Aripov* [4], *Abdelgalal O. I. Abaker* [5,*], *Azhari A. Alhag* [6] *and Alshaikh A. Shokeralla* [7]

[1]College of Sciences and Human Studies, Prince Mohammad Bin Fahd University, Khobar 34742, Saudi Arabia
[2]Statistics Department, Faculty of Science, King AbdulAziz University, Jeddah 21551, Saudi Arabia
[3]Department of Mathematics, Faculty of Science, Northern Border University, Arar, Saudi Arabia
[4]Department of Applied Mathematics and Computer Analysis, Faculty of Mathematics, NUU, Uzbekistan
[5]Department of Administrative Sciences, Applied College, Khamis Mushait, King Khalid University, Saudi Arabia
[6]Department of Mathematics and Statistics, College of Science, Taif University, Saudi Arabia
[7]Department of Mathematics, Faculty of Science, Al-Baha University, Albaha 65525, Saudi Arabia

Received: 2 Aug. 2024, Revised: 20 Nov. 2024, Accepted: : 21 Dec. 2024
Published online: 1 Mar. 2025

Abstract: Breast cancer remains a significant public health concern worldwide, prompting researchers to employ a combination of deep learning and statistical models to forecast and classify breast cancer data among women across diverse regions. This methodology facilitates the identification of mortality rate trends associated with breast cancer among females and pinpoints geographical regions with high prevalence rates of the disease. Subsequently, this enables the exploration of potential solutions. This study aims to leverage fuzzy time series and machine learning methodologies for breast cancer data prediction. The primary objective of this research is to conduct a comparative analysis between the multilayer perceptron model and the fuzzy time series model within the framework of breast cancer incidence data. This comparative analysis encompasses a spectrum of accuracy metrics to comprehensively evaluate the performance of both models. Results demonstrate the superiority of the multilayer perceptron model over fuzzy time series model, highlighting its efficacy in breast cancer prediction.

Keywords: Breast cancer, Prediction, Fuzzy time series, Multilayer perceptron, Comparison

1 Introduction

Developed countries prioritize patient data due to health importance, especially chronic diseases. In 2020, breast cancer claimed 685,000 lives globally, with 2.3 million women diagnosed. Over five years, 7.8 million women received diagnoses, making it the most prevalent cancer. Machine learning and statistical models analyze cancer prevalence, categorize data by age groups and geography, identifying vulnerable communities and common factors in breast cancer fatalities. Early detection is crucial, with various machine learning techniques available. This paper presents a machine-learning model for automated breast cancer diagnosis, utilizing Recursive Feature Elimination (RFE) and CNN techniques. Additionally, the article contrasts five algorithms: SVM, Random Forest, KNN, Logistic Regression, and Naive Bayes classifier[1]. One in three individuals faces the risk of developing cancer, a prevalent disease today. Among the critical cancers requiring early detection is breast cancer, where timely discovery significantly improves treatment outcomes. Various techniques for early detection and prediction of breast cancer are under research and application. This study aims to develop non-invasive, painless methods for early breast cancer diagnosis and prediction using data mining algorithms. Utilizing a dataset comprising measurements of frequency bandwidth, substrate dielectric constant, electric field, and tumor information, each of Weka's data mining classification algorithms was evaluated for breast cancer identification and prognosis. Results indicate that Random Forest and Straightforward CART are the two most efficient

[*] Corresponding author e-mail: aoadrees@kku.edu.sa

algorithms, achieving over 90% accuracy in detection. In this comparative study, the predictive capabilities of various data mining methods were evaluated for the diagnosis of breast cancer using a dataset derived from antenna measurements, employing a 10-fold cross-validation technique. The findings of this research indicate the feasibility of accurately detecting breast cancer tumors in patients[2]. Metastatic breast cancer stands as the leading cause of cancer-related deaths among women, primarily due to inadequate early detection measures. Through comprehensive data analysis, we have determined that utilizing blood profile data as a non-invasive machine-learning technique holds promise for early detection of breast cancer metastasis. Our findings reveal that the Decision Tree (DT) classifier outperformed the other nine algorithms evaluated, including ensemble and logistic regression models, achieving an accuracy of 83%. While the current accuracy rate may not be deemed particularly high, it has the potential for improvement by incorporating additional attributes such as serum biomarkers, vitamin D3 and B12 levels, liver and kidney function tests, and other relevant measurements. In addition to validating our hypothesis, our future endeavors will focus on deploying our web interface across hospitals with diverse specialties to establish a comprehensive cancer database with expanded attributes. By leveraging a range of statistical techniques and machine learning models, we aim to develop a precision medicine platform capable of enhancing overall survival rates and reducing healthcare costs for cancer patients, all from a single drop of blood[3].

Breast cancer ranks as the second deadliest disease globally, claiming more women's lives than any other ailment, not confined to India alone. In 2011, one in eight females in the USA was diagnosed with cancer, underscoring its widespread impact. Breast cancer can manifest benignly or malignantly due to abnormal breast cell division, leading to its development. Hence, early detection is paramount, potentially saving numerous lives and enabling successful treatment. This paper assesses the efficacy of various machine learning algorithms, including Support Vector Machine, K-Nearest Neighbor, Naive Bayes, Decision Tree, K-means, and Artificial Neural Networks, in predicting breast cancer at an early stage, utilizing the Wisconsin Diagnostic dataset [4].

Breast cancer, a leading cause of cancer mortality among women, garners significant research interest due to its prevalence and severity. This study explores the application of five classification models to breast cancer datasets to identify potential correlations for reducing breast cancer mortality risk. Comparative analysis of the models, including Decision Tree, Random Forest, Support Vector Machine, Neural Network, and Logistic Regression, highlights the Random Forest model's superior performance. The findings underscore the clinical and practical value of the developed model for real-world breast cancer diagnosis and management. Overall, this research contributes to advancing predictive modeling in breast cancer research [5].

This study employs classification techniques, including DT, SVM, RF, LR, and NN models, to predict breast cancer types based on various attributes. The outcomes of these predictions aim to minimize incorrect diagnoses and inform the development of effective treatment strategies. Two datasets were utilized in this research: the WBCD dataset comprising 699 volunteers and 11 attributes, and the BCCD dataset with 116 volunteers. Through preprocessing, 683 volunteers with 9 attributes were extracted from the raw data of the WBCD dataset, along with an index indicating the presence of malignant tumors. The classification models were evaluated based on accuracy, F-measure, and ROC curve analysis. Results favored the Random Forest (RF) model as the primary classification method in this study. These findings serve as a valuable guide for experts in accurately identifying the type of breast cancer [6].

Breast cancer stands as a significant cause of global fatalities among women, necessitating effective methodologies for data organization and analysis, particularly in the medical field. This study compares four machine learning algorithms—Decision Tree (C4.5), Naive Bayes (NB), Support Vector Machine (SVM), and k Closest Neighbors (k-NN)—on the Wisconsin Breast Cancer datasets to assess their effectiveness and efficiency in data classification. The primary aim is to evaluate accuracy, precision, sensitivity, and specificity of each algorithm, with SVM demonstrating the highest accuracy 97.13% and lowest error rate. All experiments are conducted using WEKA data in a simulated environment [7].

This study aims to compare the Multilayer Perceptron model and the Fuzzy Time Series model using breast cancer incidence data, employing various accuracy metrics for a comprehensive analysis. The research is structured into six sections: Introduction, Methodology (Sections Two and Three), Data Description (Section Four), Accuracy Measurements Results (Section Five), and Conclusion (Section Six).

2 Fuzzy Time Series

Fuzzy time series analysis is a subfield of time series forecasting that uses fuzzy logic to effectively manage uncertainty and imprecision in time series data[8]. Fuzzy time series models differ from typical time series analysis approaches by accommodating linguistic phrases or fuzzy sets instead of relying solely on precise numerical values for representation and manipulation[9]. Fuzzy time series analysis involves expressing time series data and forecasting algorithms using linguistic variables and fuzzy sets. These language variables denote qualitative descriptions, such as ”low,” ”medium,” and ”high,” instead of exact number values[10]. Fuzzy sets enable the depiction of uncertainty by giving membership

degrees to elements within a set.
The primary stages encompassed in fuzzy time series analysis comprise:
Fuzzification: This process entails transforming numerical time series data into fuzzy sets or language variables. Fuzzification methods assign each data point to one or more linguistic terms using predetermined linguistic terms or fuzzy partitions[11].
Rule Generation: In the context of fuzzy time series analysis, the process of creating forecasting rules involves examining patterns that have been discovered in past data[12]. These criteria provide the connections between the imprecise collections of past data and the imprecise collections of future forecasts. The methods for generating rules differ based on the particular fuzzy time series model employed.
Forecasting is the application of established rules to anticipate future values of a time series. These rules are based on the present and past fuzzy sets. The forecasting process entails combining the fuzzy sets based on certain criteria to provide a fuzzy prediction. This forecast is then transformed into a precise numerical forecast through the process of defuzzification.
Defuzzification refers to the procedure of transforming fuzzy sets or fuzzy forecasts into accurate numerical values. Several defuzzification approaches can be utilized, such as centroid, weighted average, and height-based algorithms [13].

3 Multilayer Perceptron

The Multilayer Perceptron (MLP) is an artificial neural network (ANN) that consists of many layers of neurons, including an input layer, one or more hidden layers, and an output layer[14]. The process of training a Multilayer Perceptron (MLP) comprises doing forward propagation to calculate predictions and then using backpropagation to adjust the weights and biases based on the mistakes in the predictions [15].
The features of the Multilayer Perceptron: **The input layer** comprises neurons that act as representations of the characteristics or input variables of the dataset[16]. Every individual neuron is associated with a specific feature, and the values that are inputted into these neurons function as inputs to the network [17].
The hidden layers: The concealed layers serve as intermediary layers situated between the input and output layers [18]. Every hidden layer is composed of several neurons, also known as nodes, which are responsible for executing computations on the inputs received from the preceding layer [19]. The neurons' activation function introduces non-linearity, allowing the network to effectively represent intricate relationships in the data.
The output layer is responsible for generating the ultimate predictions or outputs of the neural network [20]. The quantity of neurons within the output layer is contingent upon the characteristics of the given task. In binary classification tasks, a single neuron is typically responsible for expressing the likelihood of belonging to a single class. Conversely, in multi-class classification tasks, numerous neurons are employed to represent the probabilities associated with distinct classes. In regression tasks, it is customary to have a single neuron assigned to each output variable.
The activation functions are utilized to introduce non-linearity into the neural network, hence enabling it to acquire knowledge of intricate patterns within the given dataset [21]. The activation functions commonly employed in multilayer perceptrons (MLPs) encompass the sigmoid (loop) function, hyperbolic tangent (tanh) function, and rectified linear unit (ReLU) function.
The weights and biases play a crucial role in neural networks, as they regulate the strength of connections between neurons in adjacent layers [22]. Furthermore, every neuron possesses a corresponding bias term that enables it to autonomously modify its output based on the inputs In the process of training, the neural network acquires knowledge of the most advantageous values for the weights and biases to minimize the loss function.
The training: Multilayer Perceptions (MLPs) undergo training through the utilization of optimization methods, such as stochastic gradient descent (SGD) [23]. These algorithms iteratively modify the weights and biases to decrease the disparity between the anticipated outputs and the actual targets (labels or values) included in the training data.

4 The accuracy measurements

To investigate the effectiveness of the suggested models, the symmetric mean absolute percentage error (SMAPE) in equation (1), the mean absolute scaled error (MASE) in equation (2), and the mean absolute percentage error (MAPE) in equation (3) are used[24,25].

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i|}{|y_i|+|\hat{y}_i|} \tag{1}$$

$$\mathrm{MASE} = \frac{\frac{1}{n}\sum_{i=1}^{n}|e_i|}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|} \tag{2}$$

$$\mathrm{MAPE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left|\frac{\mathrm{e_i}}{\mathrm{y_i}}\right| \tag{3}$$

5 Data description

Breast cancer data among women in the United States of America were analyzed, sourced from the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention, 2021[26].

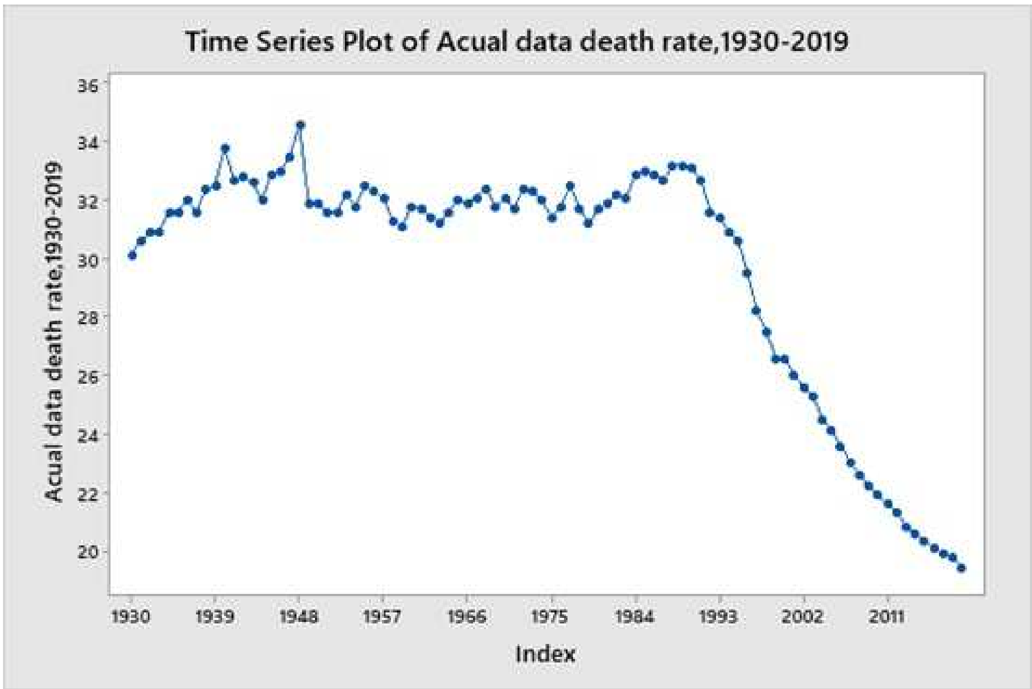


Fig. 1: Trends in death rates, 1930-2019

Table 1: The Predicted values of death rate.

Case Processing Summary			
		N	Percent
Sample	Training	63	70.00%
	Testing	27	30.00%
Valid		90	100.00%
Excluded		00	
Total		90	

The analysis of Figure 1 reveals a gradual decrease in the death rate associated with breast cancer over the years 1930 to 2019, alongside variations in mortality rates. This highlights the importance of understanding and forecasting these patterns to develop effective prevention strategies.

Figure 2 illustrates the projected estimates of death rates, generated from models constructed using the dataset. These

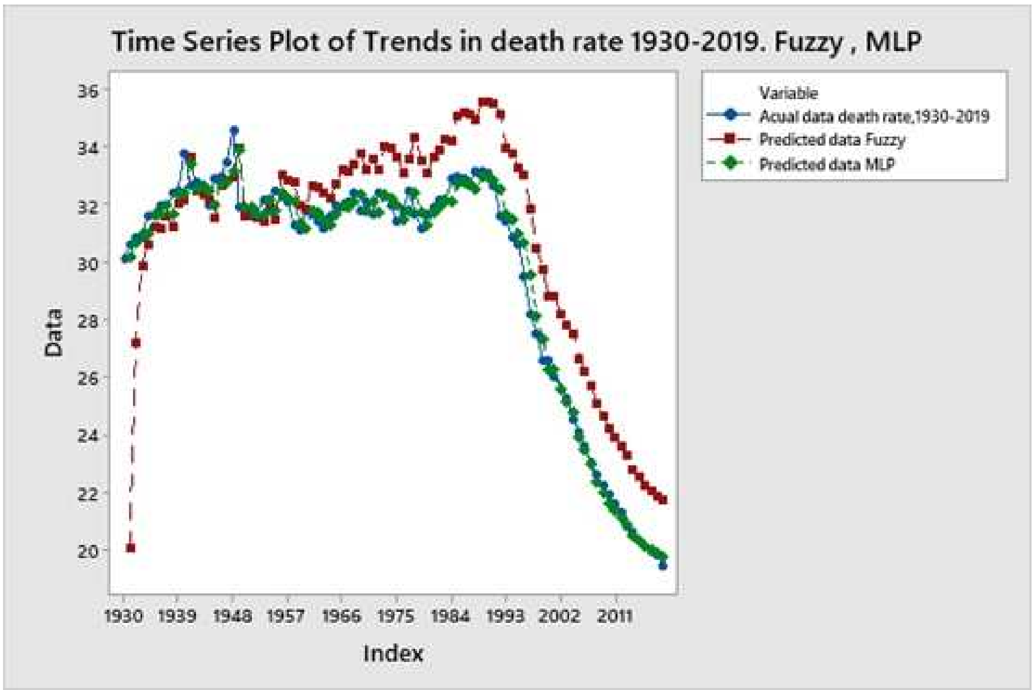


Fig. 2: The Predicted values of death rate

Table 2: The accuracy of each model

	MAPE	MASE	SMAPE
Fuzzy	0.055346441	0.36341117	0.050957629
MLP	0.006072353	0.31665608	0.006166896

projected values provide valuable insights into the future trajectory of mortality rates associated with breast cancer. Such information can be instrumental in guiding policymakers and healthcare professionals to formulate effective interventions and allocate resources appropriately. The data was partitioned into training and testing sets, with 70% allocated for training and the remaining 30% for testing. This approach ensures that the constructed models undergo adequate training and demonstratne robust generalization capabilities when applied to new data.

Table 2 presents the accuracy metrics for both the fuzzy time series (Fuzzy) and Multilayer Perceptron (MLP) models, encompassing Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), and Symmetric Mean Absolute Percentage Error (SMAPE). The findings indicate that the MLP model outperforms the fuzzy time series model, demonstrating superior predictive performance across all accuracy criteria.

6 conclusion

The results of this study underscore the superior performance of the Multilayer Perceptron (MLP) model compared to the fuzzy time series model in predicting breast cancer mortality rates. The comprehensive evaluation of accuracy metrics provides compelling evidence supporting the assertion that the MLP model is more suitable for forecasting breast cancer mortality rates. This finding carries significant implications for healthcare policy, practice, and research. By leveraging the accurate predictions generated by the MLP model, healthcare stakeholders can devise evidence-based strategies to reduce breast cancer mortality rates and improve patient outcomes. Moreover, these findings emphasize the importance of employing sophisticated modeling tools to deepen our understanding of complex healthcare challenges and facilitate informed decision-making processes. Moving forward, it is imperative to conduct further research in this area to refine predictive models, integrate additional data sources, and validate findings across diverse populations. Through the adoption of cutting-edge techniques and collaborative partnerships spanning multiple disciplines, we can advance our efforts in combating breast cancer and enhancing overall public health outcomes.

Acknowledgement

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/353/45.

References

[1] S., Bhise, S., Gadekar, A. S., Gaur, S., Bepari, D. S. A., Deepmala Kale. Breast cancer detection using machine learning techniques. Int. J. Eng. Res. Technol, 10(7), (2021).

[2] M. K., Keleş. Breast cancer prediction and detection using data mining classification algorithms: a comparative study. Tehnički vjesnik, 26(1), 149-155, (2019).

[3] M., Botlagunta M. D., Botlagunta M. B., Myneni D. Lakshmi, A. Nayyar, J. S. Gullapalli, M. A. Shah. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Scientific Reports, 13(1),(2023).

[4] T., Thomas, N., Pradhan, V. S., Dhaka. Comparative analysis to predict breast cancer using machine learning algorithms: a survey. In 2020 International Conference on Inventive Computation Technologies (ICICT), IEEE, pp. 192-196, (2020).

[5] Y., Li, Z., Chen. Performance evaluation of machine learning methods for breast cancer prediction. Appl Computer Math, 7(4), pp. 212-216, (2018).

[6] A., Aloraini. Different machine learning algorithms for breast cancer diagnosis. International Journal of Artificial Intelligence & Applications, 3(6), 21, (2012).

[7] H.,Asri, H., Mousannif, Al Moatassime, T., Noel. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, pp.1064-1069, (2016).

[8] A., Tale, A. S., Gusain, J., Baguli, R., Sheikh, A, Badar. Study of load forecasting techniques using fuzzy logic. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 6(2), pp. 512-561, (2017).

[9] E., Herrera-Viedma, I., Palomares, C. C., Li, F. J., Cabrerizo, Y., Dong, F., Chiclana, F., Herrera. Revisiting fuzzy and linguistic decision making: Scenarios and challenges for making wiser decisions in a better way. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(1), 191-208, (2020).

[10] Q., Liang, N. N., Karnik, J. M., Mendel. Connection admission control in ATM networks using survey-based type-2 fuzzy logic systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30(3), pp.329-339, (2000).

[11] J., Ares, J. A., Lara, D., Lizcano, S. Suárez. A soft computing framework for classifying time series based on fuzzy sets of events. Information Sciences, 330, pp. 125-144, (2016).

[12] P., Angelov, R., Buswell. Identification of evolving fuzzy rule-based models. IEEE Transactions on Fuzzy Systems, 10(5), pp. 667-677, (2002).

[13] J. B. Mitchell, Machine learning methods in cheminformatics. Wiley Interdisciplinary Reviews: Computational Molecular Science, 4(5), pp. 468-481, (2014).

[14] A. J., Mohammed, M. H., Arif, A. A., Ali. A multilayer perceptron artificial neural network approach for improving the accuracy of intrusion detection systems. IAES International Journal of Artificial Intelligence, 9(4), 609, (2020).

[15] A., Ahrens, C. B., Hansen, M. E., Schaffer, lassopack. Model selection and prediction with regularized regression in Stata. The Stata Journal, 20(1), pp. 176-235, (2020).

[16] F., Rossi, N., Delannay, B., Conan-Guez, M., Verleysen. Representation of functional data in neural networks. Neurocomputing, 64, pp.183-210, (2005).

[17] S. U., Jan, Y. D., Lee, I., Koo, SA distributed sensor-fault detection and diagnosis framework using machine learning. Information Sciences, 547, pp.777-796, (2021).

[18] D., Yu, J., Kang, J., Dong. Service attack improvement in wireless sensor network based on machine learning. Microprocess. Microsystems, 80, 103637, (2021).

[19] I., Idrissi, M., Azizi, O. Moussaoui. A stratified IoT deep learning based intrusion detection system. In 2022 2nd international conference on innovative research in applied science, engineering and technology IRASET, IEEE. pp. 1-8, (2022).

[20] D., Lee, K., Kim. Recurrent neural network-based hourly prediction of photovoltaic power output using meteorological information. Energies, 12(2), 215, (2019).

[21] S., Roy, K., Manna, S., R., Dubey, S. B. B., Chaudhuri. (2022). LiSHT: Non-parametric linearly scaled hyperbolic tangent activation function for neural networks. In International Conference on Computer Vision and Image Processing, Cham: Springer Nature Switzerland. pp. 462-476, (2022).

[22] O. A., Montesinos López, A., Montesinos López, J., Crossa, Fundamentals of artificial neural networks and deep learning. In Multivariate statistical machine learning methods for genomic prediction,Springer International Publishing.pp. 379-425, (2022).

[23] G. I., Kim, S., Kim, B. Jang, Classification of mathematical test questions using machine learning on datasets of learning management system questions. Plos one, 18(10), e0286989. (2023).

[24] G., Vijay, et al., Performance and Emission Prediction in a Biodiesel Engine Fuelled with Honge Methyl Ester Using RBF Neural Networks, International Journal of Mechanical and Mechatronics Engineering, 9(6), pp. 976-981, (2015).

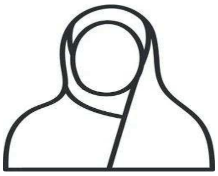
[25] S. Makridakis, , et al., Statistical and Machine Learning forecasting Methods: Concerns and Ways Forward, PloS One, 13, 3, e0194889 , (2018).

[26] https://www.cdc.gov/nchs/hus/data-finder.html

[27] https://corgis-edu.github.io/corgis/csv/can

Ibtisam Daqqa I am currently working as an Assistant professor at Prince Mohammed Bin Fahd University. My research interests lie in Linear Algebra, Differential Equations, Combinatorics, Statistics and Mathematical Education. I completed my in Ph. D. at University of South Florida in 2008. I am passionate about contributing to the field of mathematics and its applications

Abdullah M. Almarashi Professor of Statistics, King Abdulaziz University, 2023. PhD from the Department of Mathematics and Statistics, Faculty of Science, University of Strathclyde - Glasgow, UK. Head of the research group ”Inferential Statistics and its Applications” Research interests: Mathematical statistics, inferential statistics, time series analysis, regression analysis, applied statistics.Manager of the Agreements Follow-up Unit at the University Vice Presidency for Educational Affairs King Abdulaziz University, Jeddah, Saudi Arabia: Consultant to University Vice Presidency for Educational Affairs, Strategic Planning Unit, King Abdulaziz University



Mnahil. M. Bashier received her BSc in Mathematics from AL Neelain University College of mathematics in 1995, MSc in Mathematics from AL Neelain University College of mathematics in 2006, Khartoum, Sudan, and PhD in Mathematics from the Sudan University of Science and Technology (SUST) in 2016. She is currently a lecturer at the college of mathematics —Mathematics Department at Northern Border University (NBU)—Arar, Kingdom of Saudi Arabia. her current research interests include the Differential Geometrics, Liner Algebra, and Differential Equations.

M. Aripov Field of research: Mathematic modeling of nonlinear processing, Asymptotical theory of nonlinear differential equation and system, IT. The memberships of mathematical societies: AMS, GAMM, EMS, ISAAC, TWMS The projects CANDI and COCUZ focus on computational science, e-Learning, and curriculum development

Abdelgalal O. I. Abaker is associate Professor of applied Statistics at Applied College, King Khalid University, KSA. He received a PhD from Aljazera University Sudan. His research interests are in the area of Applied Statistics. His research articles have been published in international Applied Statistics journals of good repute

Azhari A. Alhag Inspirational Associate Professor in statistics dedicated to improving students learning development in mathematics and statistics sciences, resulting in achieving outstanding excellent results. Expert in course delivery and developing engaging lectures for students, increasing student satisfaction and course enjoyment. Research interests are prediction and classification using Statistical Learning Machine Learning.

Alshaikh Shokeralla is Assistant Professor of Mathematics at Al-Baha University , KSA. His research interests are in the areas of Mathematical Statistics including the mathematical modeling and simulation. He has published research articles in reputed international journals of mathematical sciences.