IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 13, April 2025

Impact Factor: 7.67

Computational Intelligence based forecasting and Investigating Life Expectancy based on Electronic Health Records

Mohit and Ritu Dagar
Department of Computer Science Engineering
Sat Kabir Institute of Technology & Management, Bahadurgarh, India
mohitbhardwajshal@gmail.com and ritudagarncce@gmail.com

Abstract: *The number of years a man anticipates to live without dying is generally referred to as his life expectancy. Regional variances, economic circumstances, sex differences, mental and physical illnesses, education, birth year, and other demographic aspects [1] are some of the elements that influence life expectancy. At present, we have electronic medical records (EMRs) which are medical records but in digital form. The life expectancy of a country population may be readily predicted with the use of EMR and cutting-edge AI algorithms. In order to estimate life expectancy, we will train a random forest model, an SVM, an XGBoost model, and a linear regression model with its different variations in this work. The original sources of the dataset were United Nations websites and the World Health Organisation (WHO) websites. The best performing model is found by using several parameters like R^2 score, Mean Squared Error (MSE) & Mean Absolute Error (MAE). During this research XGBoost model performs the best among all other Machine Learning models. Ideal forecasting helps us plan Advance Care Planning to increase our life expectancy, and it informs us what kind of therapy is needed in the early stages.*

Keywords: Life expectancy, Machine Learning Algorithm, Linear Regression, XGBoost, Random Forest, LightGBM

I. INTRODUCTION

Life expectancy plays a major role when we want to know whether the healthcare facilities are very good in any country. Life expectancy represent the local conditions of country. Country with low income and low literacy rate are more prone to death as compared to the country having maximum educated and healthy people [2]. Some country has high mortality rate due to lack of education and less healthcare facility available to human for example they don't have access to clean water supply. Such conditions i.e. income, education, healthcare facility, GDP [3] will make low life expectancy of any developed country.

Life table is used to evaluate life expectancy. It is statistical method that tells the mortality rate in population at a particular time. Life table [4] is very important because it predicts the person will die given its age. It is generally calculated for a population on a year basis. Fig. 1. represent the life expectancy of any human being for developing countries and developed countries.

By utilizing various data analytics algorithms and techniques, life expectancy can be forecasted by checking various parameters of a human being and examining them with the parameters [5] of the known person. Life expectancy plays a major role when we need to identify the mortality rates of any country and predicting life expectancy in early phase will provide a better treatment thus there is less damage is caused to patient. Also an early treatment is less expensive and there is more chance for human being to survive. Machine Learning has many great improvements in the medical field since life expectancy can be forecasted in an early stage which can help us to grant Advanced Personalized Healthcare Planning [6] for any human being.

DOI: 10.48175/IJARSCT-26027

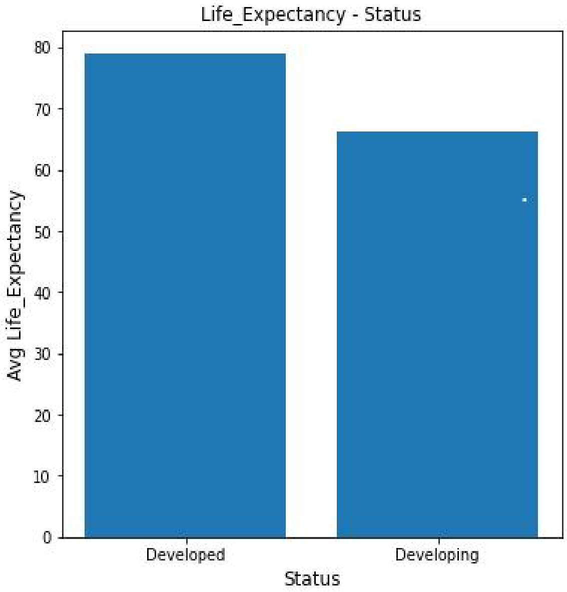

Figure 1: Life expectancy between developing and developed countries

II. LITERATURE SURVEY

We are reviewing numerous research papers and formulating a conclusion based on the findings. In 2024, the author [7] in his study shows that in terms of performance, accuracy, and efficiency, DL procedures are much better than current machine learning techniques. He evaluates the performance, accuracy, robustness, and model comparability of the taxonomy by looking at 31 cutting-edge research journal papers in the HAP system area. To assess life expectancy based on a dataset that includes economic, immunological, health, personal, and social aspects, author [8] used four machine learning algorithms namely CART, Random Forest (RF), Extra Trees, and XGBoost. The random forest model outperformed the others, according to the authors. According to him key determinants influencing Life expectancy were adult mortality, HIV/AIDS, BMI, education, and the income composition of resources based on the feature importance of this model.

The author [9] in his study utilised a wide range of regression models, such as MLR, SVR, RFR, GBR, and XGB and he finds out that the best-performing model, XGB, has the greatest R^2 value 89%. Additional refinement was accomplished by using ensemble learning strategies, particularly stacking and voting ensembles. The voting ensemble had a remarkable R^2 score of 96.7%. Author [10] explores a variety of machine learning methods to forecast life expectancy, where Extra Tree Regression yielded an adjusted R-squared value of 0.9729 demonstrating the remarkable accuracy of the model. The model found that socioeconomic, health, and environmental factors such as GDP per capita, resource composition, education levels, etc.were important predictors of life expectancy after examining a large dataset. To estimate life expectancy rates, supervised machine learning model RF and XGBoost is created [11]. eXtreme Gradient Boosting (XGBoost) algorithm applied on data from 193 UN member states, accounting for behavioural, socioeconomic, and health factors. The study's overall findings validate XGBoost as a trustworthy and effective life expectancy estimation technique. Author employ [13] two machine learning techniques used in classification for forecasting namely decision tree classification and linear regression. The results showed that the Linear Regression Classifier achieved 96% accuracy while the Decision Tree Classifier achieved 92% accuracy. Author [16] analyzed life expectancy based on various features, including immunization features like Polio, Hepatitis B etc. on the standard of living that was not previously considered. He used logistic regression, support vector machines (SVM), decision trees, and random forest regression and found that the random forest approach produced an excellent r-squared value.

DOI: 10.48175/IJARSCT-26027

The greatest significant influence on life expectancy is education [21] as seen by the 0.713054 positive correlation between the two variables life expectancy and education. Next is adult mortality, which has a correlation of -0.696390 with life expectancy indicating that life expectancy falls as adult mortality rises. Since education and adult mortality have a big influence on life expectancy, they should be considered when making life expectancy predictions. Schultz et al. [23] in his paper utilize different data science models for develop two clocks. First one is AFRAID clock and second one is FRIGHT clock. FRIGHT clock represents chronological age of mice but AFRAID clock shows forecasted life expectancy for multiple ages. Alsalem et al. [24] guess life expectancy at birth time using Boosted Decision Tree Regression. He finds out that different parameters like social factor and demographics plays very important role during forecasting

III. METHODOLOGY

A human's life expectancy may now be predicted with the highest degree of precision and accuracy due to development of data science and artificial intelligence. The input dataset, which has 22 data columns and 2938 data rows, was obtained from the WHO website. We will train our machine to predict human life expectancy using the linear regression approach and its different versions. A Random Forest regressor and XGBoost will greatly converge of dependent attributes so that our system will forecast precisely and much accurate .80% of processed input can be utilize for train our system and rest to test data so we can be sure that our machine forecast life expectancy accurately. The whole process of this research is shown in Fig. 2. In this project we develop our model using python language and various data analytics library like scikit-learn, matplot, numpy, pandas and deep learning library i.e. TensorFlow.

The dataset used for input was taken from the World Health Organization's Global Health Observatory (GHO) repository.

Since our input dataset is not cleaned and we have to filter and process our dataset to make it more suitable for applying various data analytics algorithm like removing null value, assigning useful names to columns, encoding categorical values, removing outliers. We will create a pairplot diagram shown in Fig. 3. from our input dataset to understand our data more clearly & visually.

Then we create a heatmap which shows the best parameters that helps in evaluation like expectancy. After finding correlation we break our input dataset in training and testing on basis of some percentage criterion like 80% for training and remaining data will be served for testing purpose. Lastly we applied feature scaling on our training data.

The training dataset was then subjected to a variety of machine learning algorithms in order to determine which one performed best on the testing data as well. We check accuracy of our model by measure the difference between actual and predicted value for life expectancy. It is find out that XGBoost performs best among all ML models.

Finally the life expectancy of human being is forecasted using our trained model.

Seven machine learning algorithms required for this research mentioned below:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Support Vector Regression (SVR)
- Random Forest Regressor (RFR)
- eXtreme Gradient Boosting (XGBoost)
- Light Gradient-Boosting Machine (LightGBM)

DOI: 10.48175/IJARSCT-26027

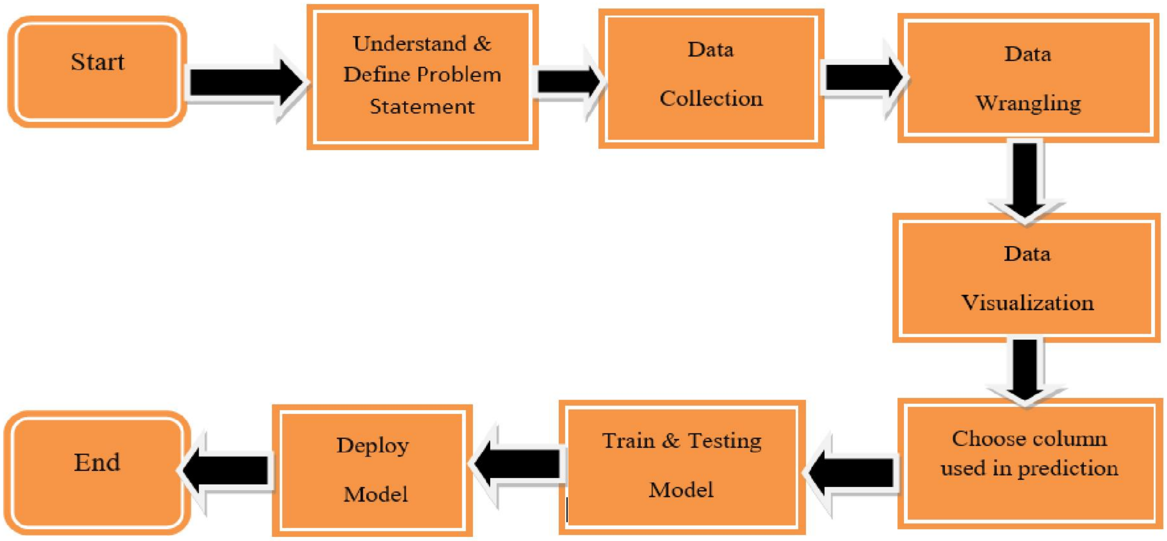

Figure 2: Block diagram of the whole research

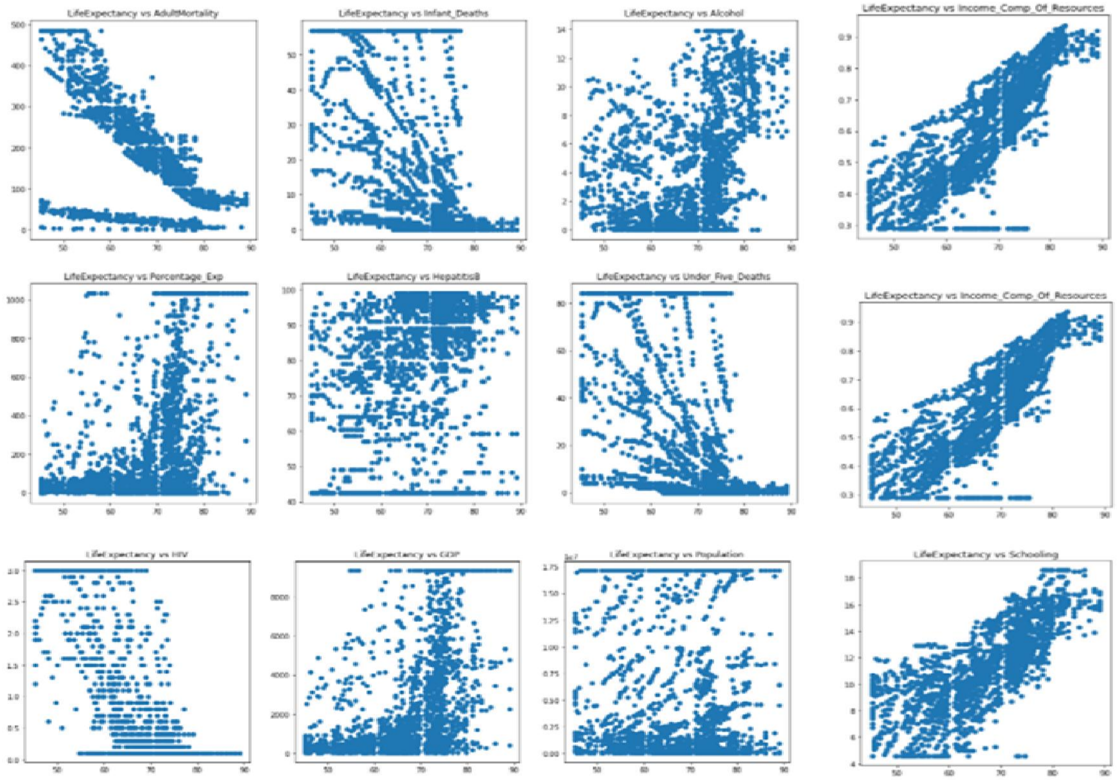

Figure 3: Pairplot diagram between life expectancy and several input features

DOI: 10.48175/IJARSCT-26027

A. Deep Learning Model

1. Linear Regression (LR): It has two variants based on the number of variable that are required for predicting dependent variable. If it uses one variable to predict desired outcome then it is known as simple Linear Regression. If it uses multiple variables then it is known as Multivariate Linear Regression. In this method our main task is guess the best line which cover maximum independent variable scatter plot.
2. Lasso Regression: LASSO regression is variant of linear regression. This method is good for those dataset that exhibits multicollinearity. It uses shrinkage method in which all data points will move towards certain point .It uses L1 Regularisation method.One problem associated with LASSO Regression is that when we have many correlated variables it preserves only one variable and other correlated variable become zero. It results lots of loss of information in data and thus generate low accuracy.
3. Ridge Regression: This method again used for data that suffers from multicollinearity like LASSO regression but it utilize L2 regularization technique .But the problem with this method it is not good for feature reduction because it does not decrease the number of variable since it does not make coefficient zero instead it minimizes it.
4. Random Forest Regression: Decision tree is computationally expensive for training the data and moreover it has problem of overfitting the data. Thus to overcome the weakness we use random forest regression tree. It is one of the best supervised learning algorithm utilizing ensemble learning in case of classification and regression problems. It is a bagging technique not a boosting technique. In ensemble method we combines results from many ML models and use that result.
5. XGBoost: XGBOOST is extreme gradient boosting technique uses ensemble Learning algorithm.It is portable, open source, support by most of language and maintained by Distributed Machine Learning Community. It utilize both hardware and software optimization to provide best result in short span of time. It has builtin cross validation method, Auto tree pruning and perform the distributed weighted Quantile Sketch algorithm for finding optimum split points. It supports parallel processing and can be used in distributed environment.

B. Performance Metrics and Loss Function

$$\text{Mean Absolute Error (MAE)} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \qquad (1)$$

$$\text{Mean Squared Error (MSE)} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \qquad (2)$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \qquad (3)$$

$$\text{Loss Function} = \text{MSE} \qquad (4)$$

Where

y_i = Actual value

$\hat{y}_i$ = Predicted value

N= Total number of observation

$\bar{y}$= Mean value

C. Heat map Chart

Fig. 4. represents the heatmap that shows the relationship between various parameters. It helps for finding the correct attribute that affect life expectancy of human being.

DOI: 10.48175/IJARSCT-26027

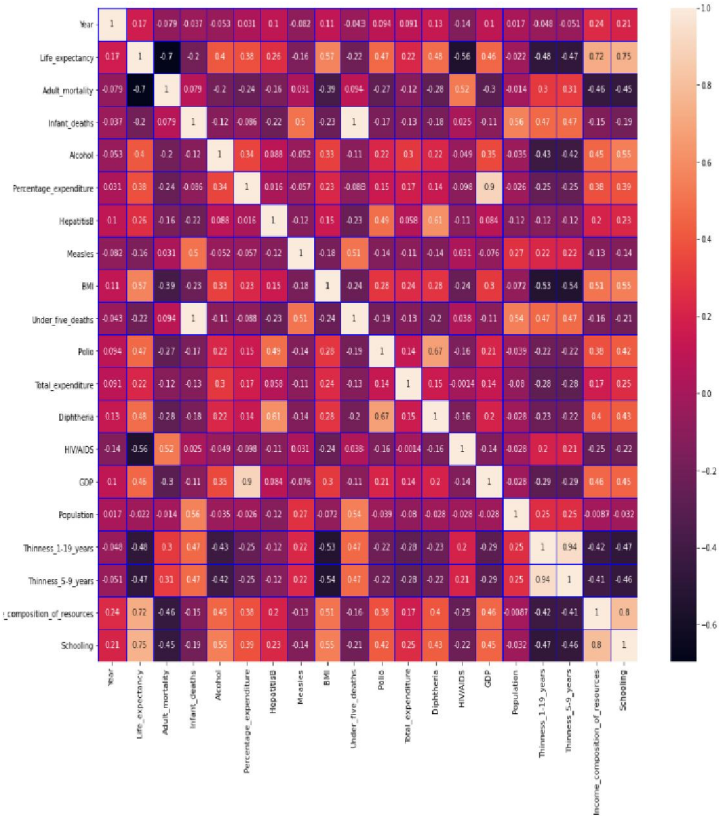

Figure 4: Heatmap for correlations between several parameter

IV. RESULT AND DISCUSSION

Different models are being compared on the basis mean squared error, mean absolute error and the R^2 score. In Fig. 5. , Fig. 6. and Fig. 7. , it is crystal clear that XGBoost perform the best on our given dataset and has much higher accuracy as compared to other model

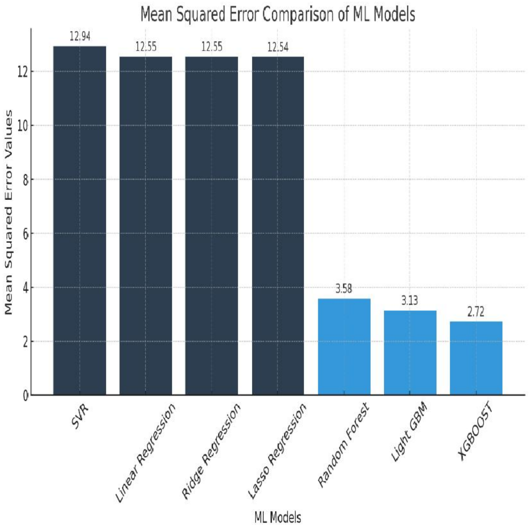

Figure 5: Examine various machine learning models via MSE

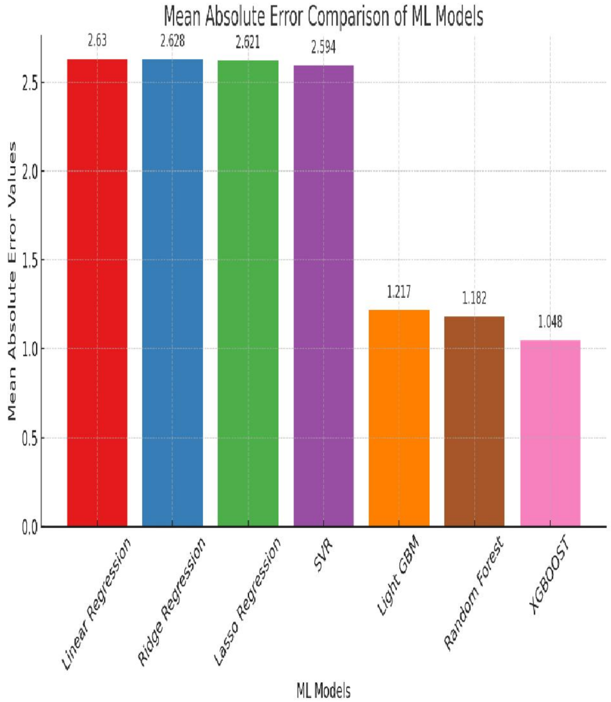

Figure 6: Examine various machine learning models via MAE

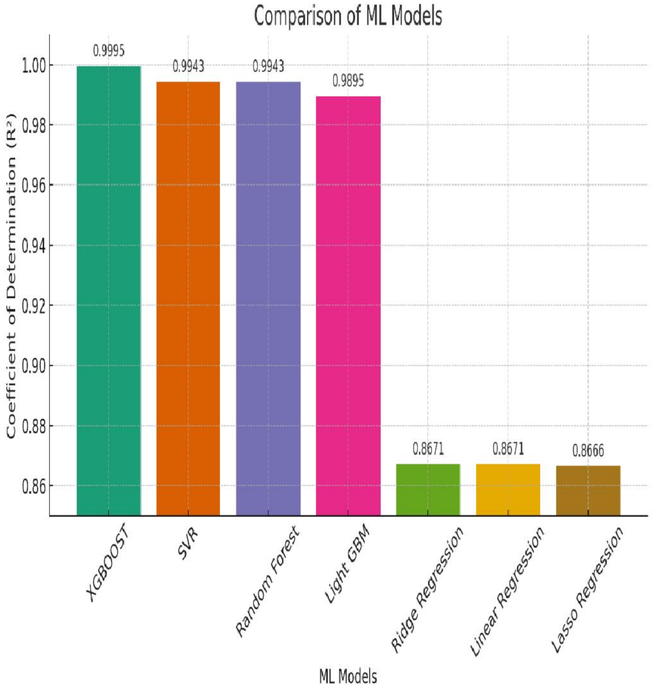

Figure 7: Examine various machine learning models via R^2 score

DOI: 10.48175/IJARSCT-26027

V. CONCLUSION

In order to anticipate and analyses human life expectancy based on several criteria, our research required a variety of data science methods, including linear regression and its different variants, RFR, XGBoost, SVR, and LightGBM. To determine which machine learning model is best for predicting life expectancy, we compared the accuracy of several algorithms. Out of all the models utilized in this study, XGBoost offers the highest accuracy, making it one of the best algorithms for estimating life expectancy. The primary motivation behind this study was to provide a faster method in predicting life expectancy for those nations lacking adequate hospital facilities.

Future Scope

In our study we use multiple machine learning algorithms to predict life expectancy. The work described in the project will be helpful in many ways. In the future, the work will be expanded in several ways:

- With a few more enhancements, it can assist us in determining the country's adult mortality rates.
- It can be integrated with IoT to create a customized healthcare system.
- More input data can be added to make model generic in nature
- A realistic framework needs to be developed that determine life expectancy in real-time environment using current abundant data and giving attention to important parameter of person.

REFERENCES

[1] J. Wang, Z. Qin, J. Hsu, and B. Zhou, “A fusion of machine learning algorithms and traditional statistical forecasting models for analyzing American healthcare expenditure,” Healthcare Analytics, vol. 5, p. 100312, Jun. 2024, doi: 10.1016/j.health.2024.100312.

[2] N. Theodorakis et al., “Integrating Machine Learning with Multi-Omics Technologies in Geroscience: Towards Personalized Medicine,” Journal of Personalized Medicine, vol. 14, no. 9, p. 931, Aug. 2024, doi: 10.3390/jpm14090931.

[3] D. U. Ozsahin, D. I. Emegano, L. R. David, A. J. Hussain, B. Uzun, and I. Ozsahin, “Global Life Expectancy Prediction Using Machine Learning Ensemble Techniques,” in 2024 17th International Conference on Development in eSystem Engineering (DeSE), Khorfakkan, United Arab Emirates: IEEE, Nov. 2024, pp. 423–427. doi: 10.1109/DeSE63988.2024.10912031.

[4] P. Narooka, A. Vashistha, P. Mitra, and G. Derashri, “Life Expectancy Prediction using Machine Learning Technique – KNN,” in 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC), Gwalior, India: IEEE, Jul. 2024, pp. 395–399. doi: 10.1109/AIC61668.2024.10730995.

[5] L.-L. Liu, H.-D. Yin, T. Xiao, L. Huang, and Y.-M. Cheng, “Dynamic prediction of landslide life expectancy using ensemble system incorporating classical prediction models and machine learning,” Geoscience Frontiers, vol. 15, no. 2, p. 101758, Mar. 2024, doi: 10.1016/j.gsf.2023.101758.

[6] K. Lakshmi, K. Deeba, V. Harave, and S. Bharti, “Life Expectancy Prediction And Diet Recommendation System for Cardiovascular and Diabetes Disease Using Machine Learning,” in 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chikkaballapur, India: IEEE, Apr. 2024, pp. 1–8. doi: 10.1109/ICKECS61492.2024.10616598.

[7] A. Alsadoon, G. Al-Naymat, and M. R. Islam, “Deep learning models for human age prediction to prevent, treat and extend life expectancy: DCPV taxonomy,” Multimedia Tools and Applications, vol. 83, no. 2, pp. 4825–4857, Jan. 2024, doi: 10.1007/s11042-023-15889-7.

[8] A. E. Ronmi, R. Prasad, and B. A. Raphael, “How can artificial intelligence and data science algorithms predict life expectancy - An empirical investigation spanning 193 countries,” International Journal of Information Management Data Insights, vol. 3, no. 1, p. 100168, Apr. 2023, doi: 10.1016/j.jjimei.2023.100168.

[9] S. Dangety, K. V. Kasulu, G. Swetha, Z. Begum, K. M. Bhashyam, and A. Lakshmanarao, “Exploring Socioeconomic Influences on Life Expectancy through Machine Learning Ensemble Regression Techniques,” in 2023

4th International Conference on Intelligent Technologies (CONIT), Bangalore, India: IEEE, Jun. 2024, pp. 1–5. doi: 10.1109/CONIT61985.2024.10626849.

[10] T. Choudhury, S. K. Bharti, M. Kumar Gourisaria, J. J. Jena, D. Kumar Behera, and A. Bandyopadhyay, "Predictive Modeling of Life Expectancy Using Machine Learning Algorithms," in 2024 Global Conference on Communications and Information Technologies (GCCIT), BANGALORE, India: IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/GCCIT63234.2024.10862085.

[11] B. A. Lipesa, E. Okango, B. O. Omolo, and E. O. Omondi, "An application of a supervised machine learning model for predicting life expectancy," SN Applied Sciences, vol. 5, no. 7, p. 189, Jul. 2023, doi: 10.1007/s42452-023-05404-w.

[12] D. Jalan, A. Tuli, V. Chaudhary, N. Sharma, and M. Rakhra, "Machine Learning Models for Life Expectancy," in 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1), Bangalore, India: IEEE, Apr. 2023, pp. 1–6. doi: 10.1109/ICAIA57370.2023.10169737.

[13] K. S. Gill, V. Anand, R. Chauhan, and M. Sharma, "Predicting Life Expectancy using Machine Learning Approach through Linear Regression and Decision Tree Classification Techniques," in 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India: IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/SMARTGENCON60755.2023.10441837.

[14] A. M. Dawoud and S. S. Abu-Naser, "Predicting Life Expectancy in Diverse Countries Using Neural Networks: Insights and Implications," International Journal of Academic Engineering Research (IJAER), vol. 7, no. 9, pp. 45–54, 2023.

[15] S. Nayak, M. Pandey, and S. S. Rautaray, "A Proposal for Life Expectancy Analysis using Machine Learning Techniques," in 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India: IEEE, Oct. 2022, pp. 1331–1335. doi: 10.1109/ICOSEC54921.2022.9951919.

[16] A. Lakshmanarao, S. A, S. R. K. T, L. G, and V. K. K, "Life Expectancy Prediction through Analysis of Immunization and HDI Factors using Machine Learning Regression Algorithms," International Journal of Online and Biomedical Engineering (iJOE), vol. 18, no. 13, pp. 73–83, Oct. 2022, doi: 10.3991/ijoe.v18i13.33315

[17] P. Chandirasekeran, S. Saravanan, S. Kannan, and V. Pattabiraman, "Analyzing Implications of Various Social Factors on Life Expectancy," National Academy Science Letters, vol. 45, no. 4, pp. 311–316, Aug. 2022, doi: 10.1007/s40009-022-01118-6.

[18] A. Brink-Kjaer et al., "Age estimation from sleep studies using deep learning predicts life expectancy," npj Digital Medicine, vol. 5, no. 1, p. 103, Jul. 2022, doi: 10.1038/s41746-022-00630-9.

[19] N. Ali, D. Srivastava, A. Tiwari, A. Pandey, A. K. Pandey, and A. Sahu, "Predicting Life Expectancy of Hepatitis B Patients using Machine Learning," in 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India: IEEE, Apr. 2022, pp. 1–4. doi: 10.1109/ICDCECE53908.2022.9793025.

[20] S. Ji, B. Lee, and M. Y. Yi, "Building life-span prediction for life cycle assessment and life cycle cost using machine learning: A big data approach," Building and Environment, vol. 205, p. 108267, Nov. 2021, doi: 10.1016/j.buildenv.2021.108267.

[21] K. Faisal, D. Alomari, H. Alasmari, H. Alghamdi, and K. Saeedi, "Life Expectancy Estimation based on Machine Learning and Structured Predictors," in Proceedings of the 3rd International Conference on Advanced Information Science and System, Sanya China: ACM, Nov. 2021, pp. 1–8. doi: 10.1145/3503047.3503122.

[22] V. Bali, D. Aggarwal, S. Singh, and A. Shukla, "Life Expectancy: Prediction & Analysis using ML," in 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India: IEEE, Sep. 2021, pp. 1–8. doi: 10.1109/ICRITO51393.2021.9596123.

[23] M. B. Schultz et al., "Age and life expectancy clocks based on machine learning analysis of mouse frailty," Nature Communications, vol. 11, no. 1, p. 4618, Sep. 2020, doi: 10.1038/s41467-020-18446-0.

[24] K. Alsalem, A. Steinmetz, N. Muhammad, D. Frierson, and M. Nashed, "Predicting Life Expectancy at Birth," in 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia: IEEE, Oct. 2020, pp. 1–6. doi: 10.1109/ICCIS49240.2020.9257630.

DOI: 10.48175/IJARSCT-26027

[25] A. I. Ebada, S. Abdelrazek and I. Elhenawy, "Applying Cloud Based Machine Learning on Biosensors Streaming Data for Health Status Prediction," 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA, Piraeus, Greece, 2020, pp. 1-8, doi: 10.1109/IISA50023.2020.9284349.

[26] A. Pandey and R. Chhikara, "Analysis of Life Expectancy using various Regression Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 209-213, doi: 10.1109/ICACCCN51052.2020.9362914.

[27] M. Beeksma, S. Verberne, A. Van Den Bosch, E. Das, I. Hendrickx, and S. Groenewoud, “Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records,” BMC Medical Informatics and Decision Making, vol. 19, no. 1, p. 36, Dec. 2019, doi: 10.1186/s12911-019-0775-2

DOI: 10.48175/IJARSCT-26027