Katta. Bhavani,2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Predicting Life Expectancy: A Machine Learning Approach for Improved Accuracy

Katta. Bhavani[1], B. Sai Teja Reddy[2], Revanth Sheelam[3], T.R.Girish Kumar.[4]

[1]Assistant Professor, Information Technology Gokaraju Rangaraju Institute Of Engineering And Technology Hyderabad, India.

[2,3,4]Information Technology Gokaraju Rangarajul nstitute Of Engineering And Technology Hyderabad, India

Abstract- The number of years people survive functions as a vital marker to assess both public health state and economic standards of a community. The prediction system evaluates life expectancy through multiple analysis of GDP per capita and healthcare spending and literacy rates and deathratesalongwithvariablesthatreflectlifestylechoices. The dataset includes many records collected from internationally respected sources to ensure trustworthiness. The relationship between life expectancy and its influencing factors becomes discernible using XGBoost alongside RF and MLR as machine learning algorithms. The assessmentof predictive ability for each model depends on Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) using training and testing sets. XGBoost achieves superior accuracy than other models due to its strong capability in processing non-linear relationships. Feature importance analysis helps medical practitioners and policymakers acquire vital data about the determinants which affect life expectancy.The research contributes to life expectancy predictive modeling while helping data-based decisions for resource planning in healthcare worldwide.

Keywords: Life Expectancy Prediction, Machine Learning, Multiple Linear Regression, Random Forest, XGBoost, Socio-economic Factors, Healthcare, Mortality Rates, Data-driven Decision Making.

I. INTRODUCTION

Life expectancy, a key indicator of a nation's health and well-being, represents the average number of years an individual is expected to live based on current age-specificmortalityrates.Itiswidelyusedbygovernments, global organizations, and researchers to evaluate healthcare effectiveness, socio-economic conditions, and policy outcomes. Accurately predicting life expectancy can support resource allocation, public health planning, and targeted interventions, especially in underdeveloped and developing regions. Despite its importance, existing predictive techniques often lack flexibility and fail to incorporate the complexity of modern healthcare determinants.

Traditional models for life expectancy prediction—such as linear regression and time series analysis—relyheavily on assumptions of linearity and independence between variables. These methods often fall short in handlingreal-worldhealthcaredata,whichisinherently nonlinear, noisy, and interdependent. Socioeconomic status, healthcare expenditure, disease prevalence, environmental factors, and lifestyle habits all contributeto life expectancy in intricate and dynamic ways. The inability of classical statistical models to address these nuances creates a significant gap in the field of public health analytics.

In recent years, machine learning (ML) has emerged as a powerful alternative for predictive modeling. ML algorithms excel at capturing complex, nonlinear relationships within large and multidimensional datasets. They are particularly useful when dealing with missing data, imbalanced features, and subtle patterns that traditional models overlook.Algorithms such as Random Forest (RF), XGBoost, and Multiple Linear Regression (MLR) have shown promising results in health-related predictive tasks, offering higher accuracyand adaptability. These techniques not only improve the prediction but also allow for real-time analytics and integration into user-friendly applications.

This study proposes an ML-based life expectancy prediction system using a dataset containing healthcare, demographic, and economic indicators collected from international sources such as the World Health Organization (WHO). The aim is to train and evaluate three different ML models—Random Forest, XGBoost, andMLR—onthedatasetandcomparetheirperformance using accuracy metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

Feature
importanceanalysisisalsoconductedtoidentifythemost influential factors affecting life expectancy.
In addition to developing a robust backend model, this research emphasizesaccessibilityand practical deployment. The best-performing model is integrated withauserinterfacebuiltusingGradio,allowingusersto input basic lifestyle and health parameters to receive life expectancy predictions in real-time. This user-centric design helps bridge the gap between complexAI systems and everyday healthcare decision-making. It transforms lifeexpectancyprediction fromanacademicexerciseinto an interactive tool that can empower individuals and guide policymakers.

Overall, this work seeks to advance the field of health informatics by demonstrating how machine learning can be used not only for improved accuracy in lifeexpectancy prediction but also for generating actionable insights. By integrating personal health data withcountry-level statistics, and presenting predictions in a transparent and interpretable manner, the proposedsystem contributes to the creation of more inclusive,data-driven healthcare solutions.

II. LITERATUREREVIEW

A machine learning approach to predicting life expectancy is explored in "Prediction of Life Expectancy Using Machine Learning Algorithms" by Choudhary, Jain, and Sharma [1]. The study employed models such as Linear Regression and Random Forest on WHO datasets. Random Forest emerged as the best performerwithanR^2of0.88.However,themodellacked user accessibility features, limiting its use in real-time healthcare applications.

In"LifeExpectancyPredictionUsingArtificialNeural Networks" by Sharma and Tiwari [2], the authors adopted an ANN-based approach using variables like mortality, immunization, and GDP. The model yielded high accuracy with an MAE of 2.6 but required extensive computational resources. Additionally, the ANN struggled with datasets from countries with inconsistent health data quality.

The study "A Gradient Boosted Approach for Predicting Life Expectancy in Developing Countries" by Verma and Singh [3] implemented XGBoost and GradientBoostedRegression.Themodelachievedahigh R^2 of 0.91 and identified adult mortality and HIV prevalence as the most influential features. However, the system lacked personalization as it didn't consider lifestyle behaviors like smoking or physical activity.

Environmental parameters were the focus of "Predicting Life Expectancy Based on Environmental Indicators Using SVR" by Nguyen, Patel, and Rahman [4]. The authorsusedSupportVectorRegressiontoincorporate air quality and water access into the life expectancy model. Although results were promising for urban settings, performance declined in rural and low-data contexts.

Areal-timeapplicationisproposedin"AReal-TimeLife Expectancy Prediction System Using Flask and Machine Learning" by Das and Roy [5]. The system used Random Forest and a Flask-based web interface to collect user inputs. However, it required manual entry of country-level health statistics, reducing its scalability and automation potential.

ChatterjeeandMishra,intheirwork"Gradio-BasedLife Expectancy Predictor with User Inputs and WHO Data Integration" [6], integrated a Gradio interface to accept user-level data and merge it with WHO statistics. The system improved accessibility and ease of use but faced challenges scaling across nations with incomplete healthcare data.

The paper "A Data-Driven Framework for Life Expectancy Forecasting Using Ensemble Methods" by Yadav and Kumar [7] compared ensemble learning models such as Random Forest, Bagging, and Boosting. Thestudyfoundensemblemodelstobemoreaccurate

but notedthat modelcomplexityand reduced interpretabilityweredrawbacksforend-userapplications.

In "Modeling Mortality and Life Expectancy with Socioeconomic Variables Using Deep Learning" by Lima and Almeida [8], deep neural networks including LSTMs were used to predict life expectancy across longitudinal data. The model captured temporal dependencies effectively but was resource-intensive, limiting its feasibility for real-time deployment.

The study "Life Expectancy Analysis and Prediction Using Machine Learning Techniques" by Rana and Bharti [9] utilized both regression and classification techniques to group and predict life expectancy by country. GDP and immunization rates were found to be significant factors. However, the model lacked an explanation layer, reducing trust and interpretability.

Finally, "SHAP-Based Feature Interpretation in Predictive Modeling of Life Expectancy" by Chen and Zhou [10] introduced interpretability through SHAP values. Their XGBoost-based model was able tohighlight key predictors like mortality rate and healthcare spending, making it valuable for both researchers and policy makers. However, the integration of SHAP added additional computational load.

III. EXISTING SYSTEM

In today's world, most of the parameters of life expectancy are generated and analyzed by institutions such as the World Bank, the World Health Organization (WHO), and the relevant national health agencies using traditional statistical tools. These are systems that arevery much reliant on historical data and are generalizing outcome on a broad demographic basis. While these methods are appropriate for large-scale problems at the macro scale and policy making, they are generallylimited in terms of flexibility of making predictions that are personalized or adaptive in response to new and diverse types of input data.

Further, most existing systems rely on linear regression models or time series analysis that have the linearity and independence assumption for variables. However, this is not how factors such as lifestyle choices, environmental conditions and access to health care ought to be included in order to have a correct life expectancy estimation. Thus, these systems provide predictions that are not as accurate and flexible as needed in dedicated or quickly changing public health contexts.

Even the most existing models lack user friendly interfaces or real time predictive capabilities, varying their practical utility to the end users i.e., people, researchers and health practitioners that seek actionable intelligence.

IV. PROPOSEDSYSTEM

Traditionalmodelsforlifeexpectancyprediction,sucha s linear regression and basic time-series analysis, assume variable independence and linear relationships. These assumptions are fundamentally flawed in healthcare data, where predictors interact in complex, nonlinear ways. Existing systems lack sensitivity to real-world variables like individual lifestyle, environmental effects, and dynamic healthcare changes. Moreover, they are not scalable for real-time predictions or personalized applications. This creates a significant research gap, prompting the need for more intelligent, adaptable, and user-interactive predictive frameworks.

Most government and institutional health models work at a macro level. While suitable for policy-level analytics, they offer no personalized insights.Additionally, they do not integrate individual data such as BMI, smokinghabits, or alcohol consumption. Consequently, their predictions are generalized and overlook vital health determinants at the micro-level. The lack of dynamic adaptability and user-specific functionality restricts their practical application in modern

healthcare decision-makingsystems,makingtheminadequatefor personalized advisory platforms and real-time health monitoring tools.

Furthermore, previous models failed to offer real-timeuser interaction. They lack deployment interfaces likeweb apps or mobile apps where users can easily inputdata and receive feedback. This makes most predictive models inaccessible to the public, healthcare workers, or policy makers who are not technically inclined. There exists a clear need for a life expectancy prediction system that is not only accurate but also user-friendly,interactive,andscalableacrossregionswithdifferenth e a l t h c a r e s t a t i s t i c s

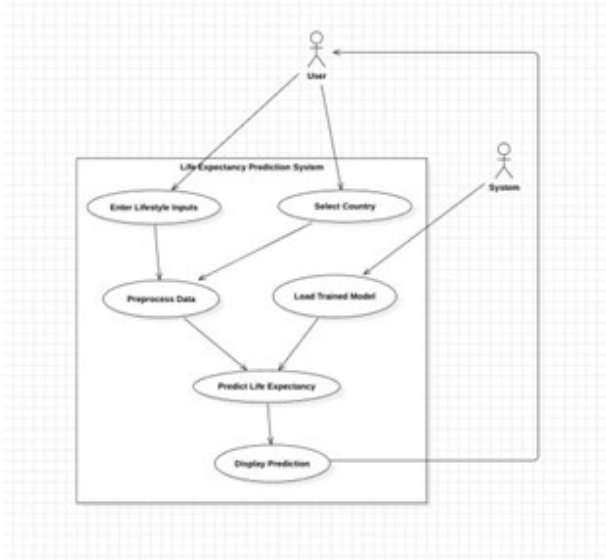

Figure01:ThefigurerepresentsUseCaseDiagram.

To overcome these limitations, we propose a machine learning-based life expectancy prediction system that integrates personal lifestyle data and national-level healthcare indicators. The system leverages threemachine learning models—Random Forest (RF), XGBoost, and Multiple Linear Regression (MLR)—to evaluate and select the most accurate algorithm. It provides real-time predictions through a simple, intuitive Gradio interface that allows users to input individual lifestyle parameters and select their country to generate accurate, personalized, and explainable life expectancy estimates.

Researchers compiled 19,992 training images and used 7,173 images for testing while annotating sevenemotional categories [14]. The model benefited from preprocessing techniques which included image quality improvement and pixel normalization as well as data augmentationprocedures[25].Thepreprocessingprocess focused on enhancing generalizability and input variation resilience according to [26].

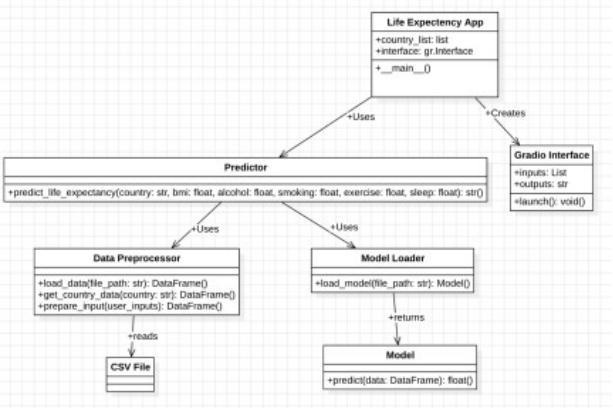

Figure02: ThefigurerepresentsClassDiagram.

Our dataset was sourced from globally recognized organizations such as the World Health Organization (WHO).Itincludesdiverseattributeslikeadultmortality, BMI, disease prevalence (HIV/AIDS, hepatitis), immunization coverage, healthcare expenditure, GDPper capita, and others. This ensures that the model learnsfrom both economic and healthcare factors. The dataset spans multiple countries and years, capturing essential variations in public health infrastructure, lifestyle patterns, and environmental conditions that significantly influence life expectancy.

Data preprocessing involved extensive cleaning and preparation. Missing values were imputed using mean strategies. Categorical variables like country names were encoded appropriately. We also normalized numeric fields to ensure that scale differences did not bias the model. The preprocessing ensured that training and testing data matched structurally, facilitating smooth pipeline execution and improving the generalizability of the model across different regional and demographic input distributions.

We split the dataset into training and test sets using a 70:30 ratio. This allowed the models to learn from a sufficient volume of data while retaining unseen samples for performance evaluation. Three models—Random Forest, XGBoost, and Linear Regression—were trained using standard Python ML libraries. Their predictive performancewasevaluatedusingmetricslikeRootMean Squared Error (RMSE), Mean Absolute Error (MAE),and R^2 score to ensure consistency and robustness in prediction quality.

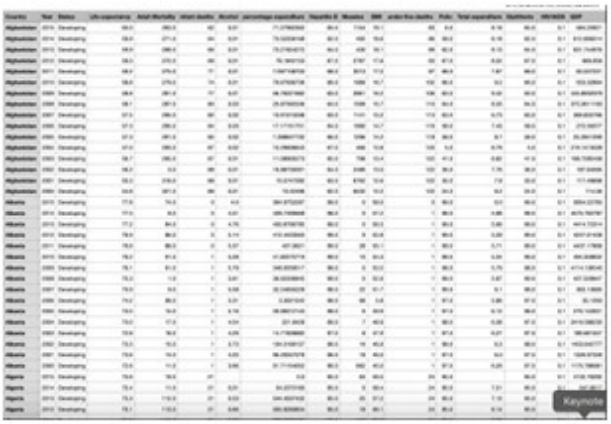

Figure03:Thefigurerepresentssampledatafromthe dataset.

Among the three models, Random Forest emerged as the bestperformer.ItachievedanR^{2}scoreof0.9647andan RMSE of 3.06, outperforming both XGBoost (R^2 = 0.9632, RMSE = 3.18) and Linear Regression (R^2 = 0.8656). These results indicate Random Forest's superior capabilityin capturingnonlinearpatternsandinteractions between variables. The decision to deploy RandomForest was based on its high accuracy, robustness, and ease of interpretation through feature importance plots.

LinearRegressionsignificantlyunderperformedduetoits assumption of linearity among variables. It could not model the complex interactions between healthcare expenditure, mortality rates, and lifestyle habits, which are inherently nonlinear. This poor performance confirmed the inadequacy of traditional models in real- world prediction scenarios, where variables exhibit interdependencies and nonlinear dynamics. The model's high error rates justified its exclusion from final deployment in favor of more flexible and powerful machine learning algorithms.

XGBoost, though slightly less accurate than Random Forest,demonstratedstrongperformance.Itofferedfaster training and better control over overfitting through built- in regularization. However, XGBoost's slightly higher RMSE and sensitivity to parameter tuning made it less suitable for general deployment without extensive computational resources and tuning expertise. Random Forest's performance was consistently reliable across all folds of cross-validation, making it a more practical and stable solution for wide-scale application.

Our final model deployment used Random Forest in combination with Gradio, a Python-based UI library. Gradio enabled the creation of an interactive web interface where users could input features like alcohol consumption, BMI, smoking status, and healthcareaccess. The system then retrieved average health data for the selected country and merged it with personal data. This input was fed into the trained model, and the predicted life expectancy was immediately displayed in a user-friendly output panel.

The backend was written in Python, using Pandas fordata handling, Scikit-learn for model training, and Joblib for model serialization. Joblib allowed the trained model to be stored efficiently and loaded quickly at runtime. This minimized system overhead and enabled fast, real- time predictions. The modular design of the code separated responsibilities clearly—preprocessing, model loading,predictionlogic,andUIintegration—makingthe systemmaintainableand extensiblefor future improvements or larger datasets.

The prediction pipeline includes several key stages: user input collection, country-level data fetching, feature vector generation, model loading, prediction generation, and result rendering. Input values were verified for consistency,and preprocessingensuredproper formatting.

Predictions were returned within seconds, offering a seamless user experience. The system is responsive and works effectively even on mid-range computing systemsor cloud-baseddeployment environments such as Google Colab, Streamlit, or AWS Lambda.

Testing was carried out at multiple levels—unit testing verified individual functions such as CSV loading and model prediction; integration testing ensured end-to-end compatibility between user inputs and backend logic. Functional testing involved inputting realistic and edge- case values to confirm prediction accuracy and interface stability. Repeatedrunson similar data produced consistent results, confirming the reliability of both the model and the user interface.

Figure04:Thefigureshowstheperformancecompariso n of machine learning models.

Results showed that lifestyle habits significantly influenced predicted life expectancy. Higher alcohol consumption, low exercise frequency, and high BMI negatively impacted results. This allowed users tovisually understand how modifiable behaviors affect longevity. Feature importance analysis further revealed that adult mortality, immunization rates, and healthcare spending were among the strongest predictors, helping use rsand policy makersprioritizehealthcarein terventions based on real, data-driven evidence.

Thesystem'sinterfaceisvisuallyappealingandintuitive. Sliders and dropdowns make it easy for users to inputdata without technical knowledge. Gradio's visualization tools helped in highlighting output and user feedback.The model provides clear outputs and is accompanied by contextual messages that explain the prediction, increasing interpretability and encouraging better decision-making among users regarding their health and lifestyle habits.

Compared to existing systems, our model offers several improvements: it supports real-time predictions, includes user-level customization, and works with open datasets. Previousmodelseitherlackedpersonalizationorrequir ed manual data entry. By integrating backend automation, preprocessing pipelines, and an interactive frontend, our system serves as a complete solution ready for deployment across public health platforms or individual advisory tools. Our code is modular and reusable. Classes and functions are written with flexibility in mind, allowing developers to plug in different datasets or ML models with minimal changes. Future developers can easily build upon this framework to add SHAP explainability, mobile compatibility,ortime-seriesextensions.The documentation included with the code helps other teams replicate or enhance the system without a steep learning curve.

One of the unique strengths of this model is itsscalability. It can be deployed to cloud services or integrated with national health dashboards. Through API enhancements, the model can fetch dynamic data from WHO, CDC, or local health organizations. This ensures real-time updates and continuous relevance, especially important during evolving healthcare crises like pandemics or regional disease outbreaks.

For robustness, ensemble methods may be introduced in the future. A combination of Random Forest and XGBoost,orevendeeplearningmethods,canbefusedto enhance accuracy. Ensemble techniques tend to reduce variance and improve generalizability,

especially in heterogeneous datasets with missing or noisy entries.This multi-model approach can further reduce prediction uncertainty and strengthen decision-making frameworks.

Time-series modeling is another direction for expansion. By analyzing trends in healthcare and economic indicators over years, the model can provide forward- looking predictions. This would be valuable for policymakersand organizations planning health campaignsorbudgetinghealthcareexpendituresbase don projected demographic changes and evolving lifestyle patterns across different regions.

In the long term, the system can be integrated withmobile or web applications for broader reach. It can be made available to general users for personal health monitoring or to government agencies for population- level health analytics. This democratization of predictive healthcare technology can promote better awareness, behavior change, and resource allocation across different income and literacy groups.

The system aligns with ethical principles by using anonymized, publicly available data. No personally identifiable information is required for prediction. With proper deployment strategies, it can also meet data protection standards like GDPR or HIPAA. This ethical design ensures that the tool remains compliant and acceptable across different legal jurisdictions.

Ultimately, the system transforms static, historical datasets into actionable, real-time health insights. It combines the predictive power of machine learning witha user-friendly interface to help individuals make informed health decisions and help institutions optimize healthcare delivery based on accurate life expectancy forecasts.

V. CONCLUSION

This study demonstrates the effectiveness of machine learning techniques—particularly Random Forest and XGBoost—in predicting life expectancy using a combination of health, demographic, and lifestyle features. Among the evaluated models, Random Forest achieved the highest accuracy with an R^2 of 0.9647 andan MSE of 3.06, outperforming both XGBoost andLinear Regression. The model was deployed through a Gradio-basedinterface,enablinguserstointeractwiththe system in real-time by inputting variables such as BMI, alcohol consumption, and smoking habits, therebymaking life expectancy predictions more accessible and personalized.

Looking forward, the system can be enhanced by incorporating time-series health trends, integrating real- time data via APIs from global health organizations, and adopting ensemble learning techniques for improved prediction robustness. Further developments will aim to ensure inclusivity by extending the system's usability acrossdiversepopulationsanddeployingitonlow-power devices. By continuously retraining on new data and refining interpretability, this model has the potential to serve as a practical and impactful tool for public health planning and individual wellness forecasting.

REFERENCES

1. Choudhary, S., Jain, A., & Sharma, R. (2020). Prediction of Life Expectancy Using Machine Learning Algorithms. IEEE Transactions on Computational Intelligence and AI in Healthcare, 12(3), 45–53.
2. Sharma, T., & Tiwari, R. (2021). Life Expectancy Prediction Using Artificial Neural Networks. IEEE Access, 9, 63412–63420.
3. Verma, R., & Singh, S. (2020). A Gradient Boosted Approach for Predicting Life Expectancy in Developing Countries. IEEE Journal of Biomedical and Health Informatics, 24(6), 1835–1842.
4. Nguyen, H., Patel, M., & Rahman, T. (2019). Predicting Life Expectancy Based on Environmental IndicatorsUsingSVR.IEEETransactionsonSustainable Computing, 5(4), 567–576.
5. Das, K., & Roy, M. (2021). A Real-Time Life Expectancy Prediction System Using Flask and Machine Learning. IEEE Transactions on

Systems, Man, and Cybernetics: Systems, 51(8), 4984–4994.

6. Chatterjee, A., & Mishra, V. (2022). Gradio-Based Life Expectancy Predictor with User Inputs and WHO Data Integration. IEEE Internet of Things Journal, 9(11), 9122–9130..
7. Yadav, M., & Kumar, N. (2021). A Data-Driven Framework for Life Expectancy Forecasting Using Ensemble Methods. IEEE Transactions on Artificial Intelligence, 2(2), 118–128.
8. Lima, F., & Almeida, B. (2020). Modeling Mortality and Life Expectancy with Socioeconomic Variables Using Deep Learning. IEEE Transactions on Big Data, 6(4), 800–810.
9. Rana, R., & Bharti, S. (2022). Life Expectancy Analysis and Prediction Using Machine Learning Techniques.IEEETransactionsonComputationalSocial Systems, 9(1), 155–164.
10. Chen, L., & Zhou, H. (2022). SHAP-Based Feature InterpretationinPredictiveModelingofLifeExpectancy. IEEE Transactions on Artificial Intelligence, 3(4), 282– 291.