# AI-Driven Life Expectancy Estimation Using Advanced Machine Learning Techniques

Farheen sulthana
Master's student, Department of Computer Science Enginnering
Shadan CollegeofEngineeringandTechnology, Pirancharu, Hyderabad, Telangana,India
Email: Farheensulthana006@gmail.com

***Abstract*— Prediction of life expectancy is an important part of public health planning and policy development, and conventional statistical methods are not always able to take into account the complex non-linear relationships between socio-economic, demographic and health factors. It is a number implemented with WHO dataset on Life Expectancy including health, demographic and socio-economic statistics on various countries. First of all, nulls were removed, and duplicates were removed. Label encoding was used for categorical variables. Regression models (RF, XGBoost, CatBoost, LightGBM) and hybrid ensembles (Voting Regressor, Stacking Regressor) were used to predict life expectancy. To assess the model performance the MSE, RMSE, MAE and R 2 score were used. The best model performance having the highest $R^2$ score was the Stacking Regressor with accuracy of 99.9% $R^2$. The feature contributions were explained using explanatory AI techniques like LIME and SHAP and a front-end application was created in Flask, with which the user could interact and make a prediction. This approach greatly enhances the accuracy and clarity of the forecast of life expectancy.**

***"Keywords— life expectancy, machine learning, CatBoost, Stacking Regressor, XGBoost, ensemble learning, explainable AI, Flask."***

## I. INTRODUCTION

Life expectancy is an important measure of the health status of a population. It is a measure of the health impact of the health care, living conditions, education, income distribution and environment [1]. Accurate estimates of life expectancy are essential for governments and health bodies to design effective public health programs, make decisions about the allocation of resources and monitor the progress of society over time. Factor longevity is more critical to understand with the changes occurring in demographic patterns and increased chronic disease burden in populations globally [2]. Advances in artificial intelligence have opened new avenues of analysing complicated health-related data and extracting useful knowledge to aid preventive and personalised health planning [3].

The traditional approaches to life expectancy estimation are based on either the statistical average, on the basis of the demographic life table, or on a linear regression based approach, the latter being the approach most frequently used [4]. These techniques, however, provide a basic understanding but typically miss nonlinear linkages, interactions among numerous health indicators, and changes among areas and people [5]. The traditional models have limited ability to predict the occurrence of [6] due to the complex relationship between factors that include immunisation coverage, prevalence of disease, mortality rates, educational attainment, economic security and health care access. Large-scale datasets dealing with various facets of health have become available in several countries, making the needs for powerful, high dimensional data analysis and handling feature types more salient.

Modern computer tools can be used to investigate health, demographic and socio-economic variables in detail which can lead to patterns and correlations being identified more effectively [8]. Transparency is particularly important in population health systems as transparency builds trust, policy relevance, and decision-making [9]. These systems can help to detect the key factors for longevity, emphasize geographical differences and inform data-informed actions to improve quality of life and mitigate health inequalities [10].

The main aim is to design an intelligent life expectancy estimation platform that utilises rich health, demographic and socio-economic information to create reliable and interpretable projections. It hopes to be used for evidence-based public health planning, to increase understanding of the determinants of longevity, and to help make decisions to improve the well-being of the people in various places, in advance.

## II. RELATED WORK

In recent years, population health analytics has shifted and become more focused on the use of predictive models for understanding longevity and health adjusted life outcomes. The initial efforts were directed towards forecasting healthy life years instead of total life years, highlighting the need to take into account functional limitations as well as survival. Nishi et al. proposed a predictive model for healthy life years without activity constraints, arguing that the health indicators and the demographic variables improve the believability of population-level health projections, when used in conjunction [11]. This book outlines the evolution from crude assessments of death to fuller evaluations of the health of populations.

Predictive models are becoming increasingly complex, and transparency and interpretability are becoming critical needs, particularly in health applications. Susnjak and Griffin pointed out the importance of explainable artificial intelligence in the setting of survival and clinical prediction, stating that transparent models promote trust and help practitioners to comprehend the reason behind the prediction [12]. The interpretability is relevant in situations such as plans for healthcare, prevention, and long-term policies. There is related work that has looked at explainable frameworks that balance accuracy and explainability to ensure outputs are actionable for decision makers.

There are some research performed to predict death from large-scale health records. Kim and Kim studied automatic prediction of death in patients with mental illness, based on electronic health records, showing the ability of ML algorithms to detect complicated health patterns, which are often missed by standard statistical approaches [13]. The findings indicate the growing importance of data-driven approaches for analyzing a wide variety of high-dimensional health data. While Nalmpatian et al. explored advanced methods for measuring mortality changes, their study underscores the importance of using current analytical

techniques to uncover long-term trends in population health and subtle shifts in mortality patterns [14].

Studies have been conducted on the impacts of some health and socio-economic factors on life expectancy. Sun used ML techniques to analyze the factors that affect the life expectancy and found that non-linear correlations between health, economic and social factors significantly affect the effectiveness of life expectancy prediction [15]. These analyses highlight the importance of having multiple features to capture the multidimensionality of longevity and population health outcomes.

Recent studies have explored models of deep learning and explainable models for LCA. Coupland et al. discussed the potential and the problems of using DL models to analyse long term health trajectories and highlighted data quality, interpretability, and generalisability across populations [16]. Their findings suggest that more complex models can capture more information, but careful design is key to ensure they remain robust and transparent for practical health applications.

Further extensions of the modelling of HLE have highlighted the need for measures of functional health. Nishi et al. performed a national cross sectional analysis and predicted healthy life years free from activity limitation and showed it was consistent between populations, indicating models could be useful for large scale public health assessment [17]. These methods demonstrate the need to consider more than survival indicators and factor in quality-of-life.

There have also been reports that the minimal features enhance usability and scalability. Kong et al. proposed a mortality prediction model that has fewer features, and demonstrated that a successful prediction can be achieved even if the data collection is less complex [18]. This approach is particularly suited for larger systems or where detailed information may not be readily available.

The discipline has also advanced through comparative studies that provide a comparison of the interpretability and performance of different prediction frameworks. Dolgopolyi et al. compared interpretable models for life expectancy estimate, and demonstrated that the transparent models can achieve competitive performance to the rest of the models and gain insightful understanding of feature importance [19]. Lastly, Rakers et al. performed a systematic review of the availability and quality of evidence for the use of predictive ML in primary care where they found the potential to be large, but methodological rigor and validation remain significant for deployment [20]. Together, these research provide a good foundation for the development of intelligent, interpretable, and scalable life expectancy estimate algorithms.

## III. Materials And Methods

The proposed system is an AI-based life expectancy estimation framework that estimates the life expectancy utilising health, demographic and socio-economic factors from huge multi-country datasets. The system employs a systematic preprocessing pipeline that enables efficient and scalable modelling, such as imputing missing values, deleting duplicate data, encoding of categorical features, etc. Advanced ensemble regression techniques are used to capture complicated nonlinear interactions and improve the reliability of prediction: RF, XGBoost, CatBoost and LightGBM, which are consistent with the current improvements in health prediction modelling [21]. The framework is enhanced further to make it more resilient and accurate using hybrid ensemble methods like Voting Regressor and Stacking Regressor that use multiple base learners to make more robust prediction over diverse populations [22]. For healthcare analytics, explainable AI methods like SHAP and LIME can be employed to decipher feature contributions and make decisions that are transparent [23]. An interactive user input interface is implemented using flask on the frontend, while predictions are created on the fly and model explanations are visualised, which makes it usable and deployable.

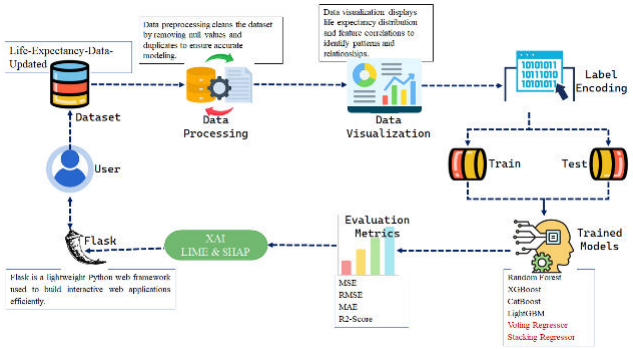



Fig.1 Proposed Architecture

The ML pipeline presented in Fig.1 starts with a user submitting a dataset of life expectancy. This involves data pre-processing, visualisation, label encoding and train test split. Various models are tried such as RF, XGBoost and various measures such as R2, MSE etc. Finally, a web app with Flask is developed for transparency with the use of XAI tools (LIME/SHAP).

*i) Dataset Collection:*

Life-Expectancy-Data-Updated", which is presented in the dataset overview in Fig.2, contains 2,864 rows and 21 columns and provides comprehensive health picture of the world. All data is architecturally tidy, and there are no NULLs for any feature. It has health measures like “Infant_deaths”, “Adult_mortality”, “Alcohol_consumption”, and vaccine coverage for “Hepatitis_B” and “Polio” as well as a few category identifiers like “Country” and “Region”. “GDP_per_capita”, “Population_mln” and “Schooling” are economic and demographic factors. It has physical measurements such as “BMI”, “Thinness” and others, in addition to the goal variable, “Life_expectancy”. Multivariate structure enables us to explore in detail the relationships between SES, public health actions and longevity in various regions of the world.



Fig.2 Dataset Collection

*ii) Pre-processing:*

Data pre-processing is a technique used to preprocess data to ensure that the data is of good quality and reliable. Corrects inconsistencies, incompleteness and redundancy that could impact model performance. This is a systematic process of cleaning and converting the data into a consistent and ordered set of data for good learning and minimal noise. Predictive

models are more robust and generalisable when they are preprocessed.

*a) Data Processing:* Data Processing is a phase that involves cleaning and arranging the data into accurate and consistent dataset before developing a model. Values of null or missing are identified and removed from data to prevent the data from being used for any biased learning and inaccurate prediction, particularly when the missing information cannot be accurately imputed. Duplicates need to be removed as well because they can distort statistical distributions and have a disproportionate impact on model results. The data is reviewed to identify any duplicated records based on the important characteristics and only unique copies are kept. This technique provides the maximum data integrity, minimum data duplication and faster performance. The process of cleaning the data ensures that the pattern in which the data is fed into the ML algorithms is consistent, thus ensuring convergence and stability of the algorithms. To sum up, efficient data processing means that the data set is representative of the real-world scenario, has consistent interaction with features, and provides a solid foundation from downstream tasks such as visualisation, encoding, and predicting prediction on other populations.

*b) Data Visualization:* Visualising data to identify patterns, trends, and relationships in the data; provides intuitive insights to aid in modelling decisions. Life expectancy . What is the range , the average , the variation in life expectancy among populations ? Plotting this distribution helps to identify skewness, outliers and probable data imbalance which may affect model selection and evaluation. To explore the correlations between health, demographic and socio-economic characteristics, a correlation matrix was used. The correlation coefficients are shown in this matrix, and strongly correlated features, redundant variables and potential problems of multicollinearity can be identified. These are useful for gaining insight into the qualities and which ones to select — the models learn important relationships and not noise. Summarizing, visualisation can be used as a way of analysing the data, identifying the pre-processing, modelling methods and present a holistic perspective of the relationships between various factors influencing the life expectancy .

*c) Label Encoding:* Label encoding is a method that is utilised to convert category variables into numerical values that may be understood by ML algorithms. Numerical models cannot easily be understood by many health, demographic and socio-economic attributes which are categorical, such as region, gender or education level. Label encoding is used to assign a unique integer value to each category, ensuring data consistency. The transformation enables ease of use of categorical variables in regression and ensemble based models. The encoding is consistent and applied equally to the training and testing data, to avoid inconsistencies. Label encoding helps reduce the difficulty of the data, ensures that algorithms are followed and supports efficient learning without altering the information embedded in the data. The encoded characteristics so help in capturing links between category parameters and life expectancy outcomes.

*iii) Train & Test:*

The train-test split is an important phase in developing effective prediction models as it helps to objectively evaluate the performance of the model. In this case, the data set is partitioned into an 80:20 split, with 80% of the data set taken to be the training set, and 20% of the data set taken to be the testing set. The training set is utilized to find out the underlying patterns, correlations and interactions between features and life expectancy. During this phase the model parameters are tuned in order to reduce the prediction error. The testing set is NOT used to train the model - it is used to determine the “generalisation” of the model to new data. This split can help identify overfitting and gives unbiased estimation of the actual performance. An 80-20 split is a good compromise between providing a sufficiently large training set to capture complex patterns, and a sufficiently large testing set to validate complex performance.

*iv) Algorithms:*

RF are a method of ensemble learning. During a model's training phase, it constructs many decision tree models. Combination of many trees to obtain more accurate and stable results, which helps to prevent overfitting and enhance generalisation. This can be particularly helpful when you have regression and classification problems with complex non-linear relationships. RF takes a random subset of data and a random subset of features to create a variety of trees such that each tree learns various features of the data. In the life expectancy estimation system, Random Forest models the relationships of health, demographic and socio-economic factors to predict life expectancy. This is a good method to work with high dimensional data, provide feature importance and deal with missing values, and is a powerful approach to gaining valid predictions for multi-country data.

XGBoost is an implementation of gradient boosting which incrementally adds an additive model whose performance is maximized through gradient-based minimization of a loss function. It is more concerned with the erroneous errors of the previous models and increases the accuracy of the predicted errors. XGBoost makes use of regularisation to stop over fitting the model and can be executed in parallel, making it very efficient and scalable for large datasets. In the prediction of life expectancy, XGBoost can learn complex nonlinear relationships among many features, improving the accuracy of the predictions. The missing value management, tree trimming algorithms and feature weighting are built-in, allowing for a comprehensive model of multi-country health and demographic data. This is useful for generating credible and meaningful life expectancy forecasts for various populations.

CatBoost is a gradient boosting technique that is able to deal with categorical data without requiring a lot of preprocessing. It is based on ordered boosting to prevent prediction bias and overfitting. It does a pretty good job for regression problems consisting of numerical and categorical variables. CatBoost handles nonlinear relationships between socio-economic, demographic and health factors in the estimation of life expectancy with high accuracy. It can directly process categorical variables, which decreases the need for preprocessing and increases the interpretability. To find out what features are most important for the predictions, CatBoost also employ regularisation and permutation importance analysis. It provides reliable projections of life expectancy at the population level in several countries, by efficiently modelling relationships and patterns in complex data.

LightGBM is a distributed and efficient gradient boosting framework based on histogram-based method and leaf-wise

growing trees. Computationally efficient, it can handle large numbers of data points and high-dimensional data. LightGBM decreases memory use and speeds up training compared to conventional boosting. It forecasts suitable values for life expectancy estimates, by modeling the correlation between health, demographic and socio-economic indicators. LightGBM can handle categorical features, parallel computing and regularisation to guarantee strong generalisation. It has the ability to learn from a large number of datasets with diverse patterns, making it suitable for a wide range of datasets from multiple countries with numerous linked elements to predict life expectancy in a robust and scalable manner.

The Voting Regressor is an ensemble approach, in which multiple predictions from multiple base regressor models are aggregated together. It is the combination of predictions from multiple algorithms (averaging/weighting) to minimize the bias and variation and to enhance the overall performance of prediction. VR improves life expectancy predictions by using the predictions of other regressors like RF, XGBoost, CatBoost and LightGBM. The advantage of each model is exploited and the errors of one model are corrected by the other models; this is the technique used. Voting Regressor performs best when the base models have similar performance, providing robust and stable estimates of life expectancy across different sets of multi-country data.

Stacking Regressor is an ensemble learning technique, which uses a meta-model to integrate the outputs of numerous base models to increase the forecast accuracy. The base models are trained separately and their predictions are used as inputs to the meta-model that learns to optimise the final predictions. The RF, XGBoost, CatBoost and LightGBM models are used to predict life expectancy and their predictions are combined to form life expectancy predictions using the Stacking Regressor then the output is improved and accuracy increased by the use of a meta-learner. This technique captures complex interactions between features, with each base model's capabilities. This aids to lessen the bias as well as variance. It is superior to the state-of-the-art in the task of making accurate, interpretable prognoses of life expectancy on multi-country datasets.

## IV. RESULTS AND DISCUSSIONS

MSE: The MSE is a measure of the error of statistical models. It is the average of the squared errors, that is, the mean squared difference between the estimated values and the real one. If the model is correct, then the mse will be equal to zero. The more the model error, the more the model error value. The mean square error is often referred to as the MSD.

$$"MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (1)"$$

RMSE: RMSE is the mean error between the values estimated by a statistical model and the actual values. From a mathematical point of view, it is the standard deviation of the residues. The residuals are the distances from the data points to the regression line.

$$"RMSE = \sqrt{\frac{\sum_{i=1}^{n}||y(i) - \hat{y}(i)||^2}{N}} \qquad (2)"$$

MAE: How different your measurements are from the actual value. It is the difference between the measured value and the "true" value. The absolute error of a scale reading of 90 lbs when you know you are 89 lbs is 90 lbs – 89 lbs = 1 lbs.

$$"MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (3)"$$

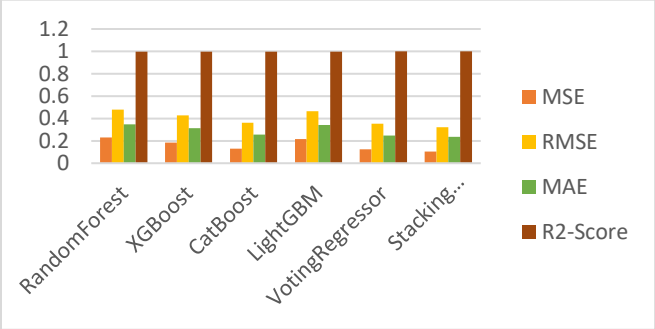
R2 Score: The sum of the residuals squared is the R2 Score Sum Squared Regression. Total Sum of Squares is the sum of the distance the data is away from the mean all squared.

$$"R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad (4)"$$

"Table.1 Performance Evaluation"

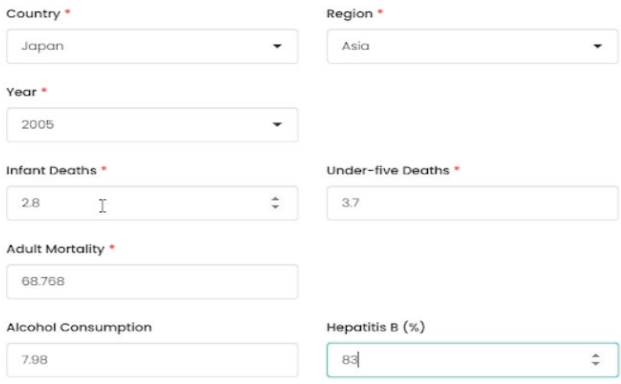
| ML Model | MSE | RMSE | MAE | R2-Score |
|---|---|---|---|---|
| RandomForest | 0.229 | 0.479 | 0.347 | 0.997 |
| XGBoost | 0.183 | 0.428 | 0.312 | 0.998 |
| CatBoost | 0.130 | 0.361 | 0.257 | 0.998 |
| LightGBM | 0.216 | 0.464 | 0.343 | 0.997 |
| VotingRegressor | 0.124 | 0.353 | 0.248 | 0.999 |
| Stacking Regressor | 0.104 | 0.322 | 0.237 | 0.999 |

A comparison between the regression models is done based on the MSE, RMSE, MAE and R2-score in Table 1. These results indicate that the ensemble is more successful.



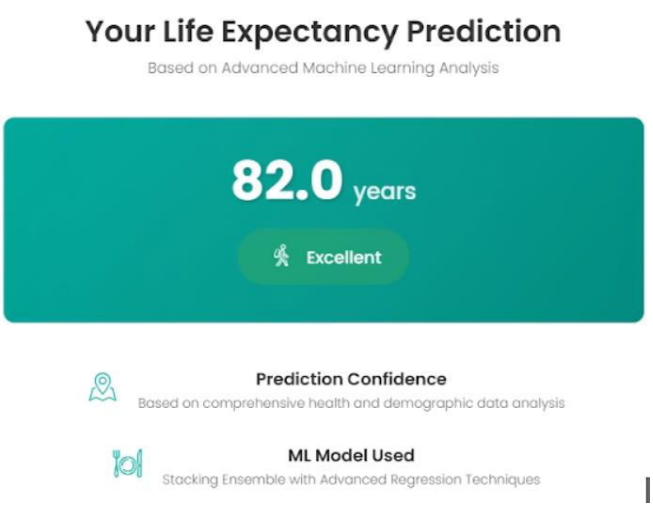

"Fig.3 Comparison Graph"

Fig.3 models' performance in predicting life expectancy using various error measures and high R2- Scores.



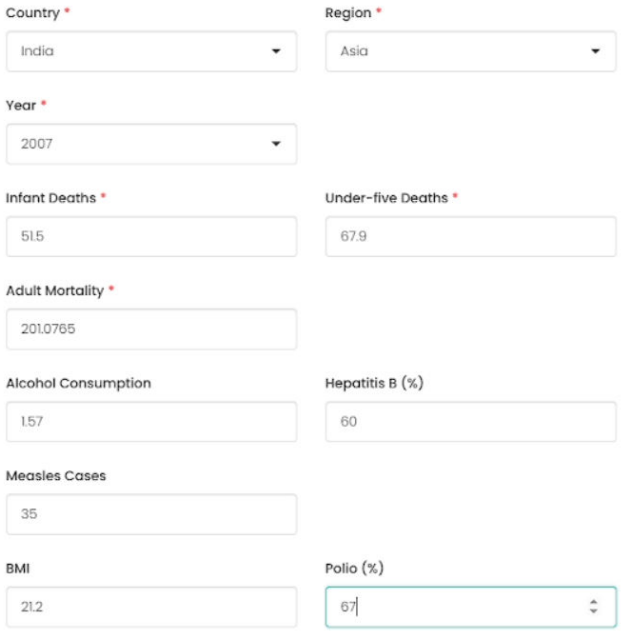
"Fig.4 Upload Input"

The health and economic data entered in the Flask application using the user interface is displayed in Fig.4.



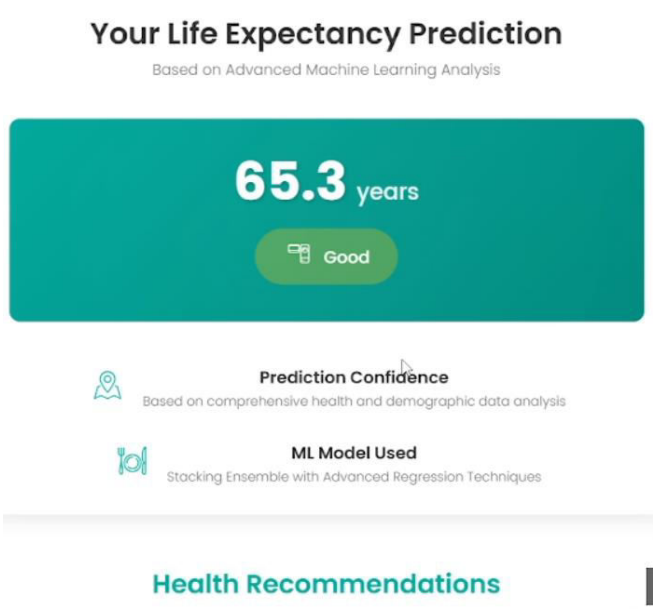

"Fig.5 Predict Result"

The last life expectancy forecast using Stacking Ensemble model is shown in Fig.5. The predicted life expectancy = 82.0 years.




"Fig.6 Upload Input Data"

The input interface for entering demographic and health data to the prediction model is presented in Fig.6.




"Fig.7 Final Outcome"

The advanced model gives a prediction of 65.3 years, shown in Fig.7, and classified as 'Good'.

## V. CONCLUSION

The adoption has resulted in a highly accurate life expectancy estimation, thanks to the advanced ML techniques applied to multi-country health, demographic and socio-economic data. The data was cleaned and preprocessed, with missing values imputed, duplicate entries removed, and categorical features encoded, to ensure the data was consistent and suitable for building powerful models. The model with the best predictive performance among the algorithms implemented was the Stacking Regressor with an $R^2$ of 99.9%, a MSE of 0.104, RMSE of 0.322 and an MAE of 0.237, showing its ability to effectively model complex nonlinear relationships. XAI techniques, such as LIME and SHAP, were used to visualize the most important features to understand the predictions and find key factors that affect life expectancy. To improve the practical usability and accessibility of the model, a Flask-based frontend was added and the user can enter the health and demographic parameters interactively, and get the prediction results in real time. In general, this makes life expectancy estimation more accurate, transparent, and interactive, providing policymakers, healthcare providers and researchers with a useful tool to inform decisions by predictive life expectancy.

Future research could include extending the data set to include more recent and varied country-specific health, economic and environmental data to increase the generalizability of the framework to other populations. Analysis of time-series and longitudinal data could be used to enhance the prediction of life expectancy over time. Non-linear and sequential data structures, like LSTM or Transformer based models, might be more appropriate for deep learning architectures. Further explorations can be made on advanced ensemble methods and hyperparameter optimisation methods to create hybrid ensemble strategies. Other approaches like as Integrated Gradients or Counterfactual Analysis can be used to better interpret the

model's conclusions within the explainable AI framework. Finally, it could create the opportunity to integrate real-time data streams with policy scenarios for proactive healthcare planning and making strategic choices.

## REFERENCES


[1] Taiwo, K. A. (2025). AI in population health: Scaling preventive models for age-related diseases in the United States. International Journal Of Science And Research Archive, 16(01), 1240-1260.

[2] Xu, K. (2025, April). AI-Driven Personalized Fall Prevention for Older Adults. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 28, pp. 29610-29612).

[3] Wilczok, D. (2025). Deep learning and generative artificial intelligence in aging research and healthy longevity medicine. Aging (Albany NY), 17(1), 251.

[4] Valavanidis, A. Artificial Intelligence (AI) Can Revolutionize Dementia Diagnosis and Treatment through Advanced Technologies.

[5] Das, A., Uddin, M. M., & Karim, M. R. (2025). Predicting life expectancy using machine learning techniques. International Journal of Statistical Sciences, 25(1), 55–70.

[6] Yang, J., Li, Z., Wu, M., Zhang, Y., Zhang, B., & Shi, H. (2025). Identifying subjective life expectancy risk factors in physically active and inactive middle-aged and older adults using machine learning models. BMC Public Health, 25, Article 3506.

[7] Athavale, S. K., & Cuturrufo, A. (2025). Modeling life expectancy with interpretable machine learning approaches. NHSJS.

[8] Chen, S., Yu, J., Chamouni, S., & Wang, Y. (2024). Integrating machine learning and artificial intelligence in life-course epidemiology: Pathways to innovative public health solutions. BMC Medicine, 22, Article 354.

[9] Malanin, V. (2025). Machine learning algorithms in personalized life expectancy prediction: A systematic review. International Journal of Science Engineering and Management.

[10] Dolgopolyi, R., Amaslidou, I., & Margaritou, A. (2025). Interpretable machine learning for life expectancy prediction: A comparative study. arXiv.

[11] Nishi, M., Nagamitsu, R., & Matoba, S. (2023). Development of a prediction model for healthy life years without activity limitations. JMIR Public Health and Surveillance, 9, e46634.

[12] Susnjak, T., & Griffin, E. (2023). Towards clinical prediction with transparency: Explainable AI in survival modeling. arXiv.

[13] Kim, S., & Kim, S. (2023). Automatic prediction of mortality in patients with mental illness using EHR. arXiv.

[14] Nalmpatian, A., Heumann, C., & Pilz, S. (2023). Advanced techniques in mortality trend estimation. arXiv.

[15] Sun, J. (2024). Analyzing key health factors influencing life expectancy using machine learning. Applied and Computational Engineering, 86, Article 1601.

[16] Coupland, H., Scheidwasser, N., Katsiferis, A., et al. (2025). Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis. BMC Public Health, 25, 1520.

[17] Nishi, M., Nagamitsu, R., & Matoba, S. (2023). Development of a prediction model for healthy life years without activity limitations: National cross sectional study. JMIR Public Health and Surveillance, 9, e46634.

[18] Kong, T. H., et al. (2025). Machine learning based mortality prediction model using minimal features to assist end of life decision making. Cureus, 17(11), e95874. PMID: 41341340

[19] Dolgopolyi, R., Amaslidou, I., & Margaritou, A. (2025). Interpretable machine learning for life expectancy prediction: A comparative study. arXiv.

[20] Rakers, M. M., et al. (2024). Availability of evidence for predictive machine learning algorithms in primary care: A systematic review. JAMA Network Open, 7, e32990.

[21] Nishi, M., & Matoba, S. (2023). Healthy life years prediction modeling. JMIR Public Health and Surveillance, 9, e46634.

[22] Integrating machine learning and cancer mortality prediction. (2025). Vershinina, O., Sabbatinelli, J., Bonfigli, A. R., et al. arXiv.

[23] Stahl, D. (2024). New horizons in prediction modelling using machine learning in older people's healthcare research. Age and Ageing.

[24] A machine learning approach to predict biological age and its longitudinal drivers. (2025). Dunbayeva, N., Li, Y., Xie, Y., & Razzak, I. arXiv.

[25] Machine Learning enabled all cause mortality and cohort risk models. (2024). Qi Yu, L. Zhang, Q. Ma, et al. Heliyon, e36878.