ZKG INTERNATIONAL ISSN: 2366-1313

Supervised Machine Learning Algorithms Based Cancer Death Cases Forecasting

[1] Velagana Akhilandeswari, [2] S. Aruna

[1] MCA Student, Dept. Of MCA, Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapur, Andhra Pradesh 534280,

akhilavelagana@gmail.com

[2,] Assistant Professor, Dept. Of MCA, Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapur, Andhra Pradesh 534280,

Abstract: *In India, like within the relaxation of the world, most cancers are a major killer. This research objective is to be expecting cancer mortality in India, the usage of supervised gadget mastering strategies. Cancer mortality quotes in India between 1990 and 2017 are provided by using age group, gender, and region the use of information from the Global Burden of Disease Study. We hire 3 distinct supervised studying algorithms—linear regression, choice tree regression, and random woodland regression—after acting information pre-processing, which incorporates missing price imputation and characteristic engineering. Using a spread of standards, we examine the effectiveness of those fashions and finish that the random wooded area regression model is advanced to the alternative. The scope of research is offer an extended-time period prediction of cancer mortality in India the usage of the satisfactory version so it'll help health branch to work on it. Our research has implications for policymakers and healthcare companies in India, wherein it may inform efforts to lessen cancer prices and enhance cancer care.*

Keywords— *Indian Cancer, Cancer Deaths, Forecasting, Supervised Machine Learning, Linear Regression, Decision Tree Regression, Random Forest Regression, Polynomial Regression, Evaluation Metrics.*

I. INTRODUCTION

In India, like in the relaxation of the globe, cancer is a main killer. Population boom, age, adjustments in way of life, and environmental elements are all contributing to a rising tide of cancer in India. National Cancer Registry Program projections placed the wide variety of latest most cancers diagnoses in India at 1.39 million in 2025, up from an expected 1.16 million in 2018. Predicting cancer fatalities in India is crucial for making plans and allocating resources for most cancers prevention, early detection, and

ISSN: 2366-1313

remedy.

Cancer prediction and analysis are most effective two regions where device mastering has showed promising enhancements in current years. Predictions of mortality, morbidity, and disease occurrence have all visible massive utility of supervised machine mastering methods. To be expecting cancer mortality in India, we gift a supervised gadget learning method. Cancer mortality prices in India between 1990 and 2017 are furnished by way of age institution, gender, and location using information from the Global Burden of Disease Study.

The purpose of these studies is to evaluate and evaluation the efficacy of three supervised learning algorithms for predicting most cancers mortality in India: linear regression, choice tree regression, and random forest regression. To expect most cancers mortality in India over the following five years, we examine every version the use of a ramification of standards together with accuracy, precision, consider, and F1 score earlier than choosing the one that performs the fine. This study has the capability to aid choice-makers and healthcare vendors in India of their efforts to higher prevent and cure cancer.

The remainder of the paper is organised as follows: In Section 2, we assessment the literature on medical machine studying and cancer prognostic prediction. Section three carries the specifics of the methodology used. Section 4 details the polynomial regression strategies used within the proposed framework. Section five provides our have a look act's conclusions, which includes the consequences of our version assessment and our forecasts. The study concludes with a discussion of the constraints of the paintings and recommendations for in addition studies.

II LITERATURE REVIEW

The costs of both new cases and deaths from cancer are anticipated to upward thrust inside the future years, making it a major public fitness concern. Numerous research have attempted to forecast most cancers incidence and dying prices the use of statistical models and device getting to know algorithms in order to better recognize and count on the burden of cancer. The reason of this literature assessment is to synthesise the outcomes of contemporary studies into the trouble of estimating most cancers occurrence and death rates across geographic areas and medical paradigms.

In [1] K. Sathishkumar et.Al used facts from the National Cancer Registry Programmed to be expecting cancer prevalence fees for 2022 and 2025 in India. According to the studies, the cancer

ISSN: 2366-1313

burden in India is expected to raise from an expected 1.4 million new cases in 2022 to one.7 million new cases with the aid of 2025.

In [2] C. Tudor et.Al in Romania evolved a innovative technique for modelling and predicting cancer incidence and dying charges. The research, which trusted Google Trends facts, found that internet queries had a giant dating with most cancers costs and fatality rates. The model predicted an upward thrust in most cancers prices and deaths in Romania among 2015 and 2025.

In [3] R. Gupta et. Al employed regression fashions to forecast automobile prices. While this research isn't right now applicable to estimating most cancers costs, it does display how regression models may be used on this context.

In [4] M. Dalmartello et.Al researched the projected most cancers death fees in Europe in 2022, with an emphasis on ovarian cancer. The look act's projection of an growth in fatalities from ovarian most cancers in 2022 in comparison to 2021 emphasises the want of detecting the disease early and supplying appropriate treatment.

In [5] B. Trächsel et.Al the studies, which depended on data from a most cancers registry, determined that the incidence of most cancers become anticipated to rise and the mortality charge to rise as well by using the yr 2025. The research also stressed the importance of cancer prevention programmes in Switzerland for lowering the United States of America's cancer price.

In [6] K. W. Jung et.Al used a statistical model to challenge cancer incidence and death prices in Korea till the yr 2022. According to the survey, cancer fees are expected to upward push dramatically from the preceding 12 months. The findings of the research also confused the want of imposing efficient most cancers control measures in Korea.

In [7] B. Sekeroglu and K. Tuncal et.Al to forecast cancer quotes during Europe. The research revealed a high degree of accuracy in the gadget learning fashions for forecasting most cancers incidence quotes, and that they utilised information from the World Health Organization. The studies showed that gadget studying fashions may also help in cancer prevention and control.

In [8] Shaikh et al. Used regression fashions and time collection forecasting to look at and make predictions on COVID-19. Predictions of COVID-19 times had been made the use of two time series forecasting strategies (ARIMA and Prophet) and 3 regression fashions (linear, ridge, and Lasso). When as compared to

ISSN: 2366-1313

the other fashions, the findings showed that the Prophet algorithm supplied the most accurate predictions.

In [9] Rahib et al. In studies modelled and predicted cancer prevalence and mortality costs using facts from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) programme. Several varieties of most cancers, with breast cancer being the maximum not unusual, confirmed an upward fashion in occurrence prices, as did lung cancer.

In [10] Jain et al. Predictions of breast most cancers incidence had been made the usage of 4 one of a kind system gaining knowledge of techniques (logistic regression, choice tree, K-NN, and help vector machine). The findings tested that the SVM algorithm furnished the maximum correct predictions as compared to the other fashions tested.

In [11] J and Jakka et.Al the use of system learning models [11]. The research predicted the COVID-19 instances the usage of numerous device studying fashions, inclusive of SVM, decision tree, and random woodland. The findings showed that the SVM model supplied the most accurate predictions compared to the other fashions.

In [12] E.-D. J Med Discov and Xie et.Al the research analysed and expected cancer prevalence quotes using the Box- Jenkins ARIMA version. Incidence quotes of many most cancers types had been efficiently expected the usage of the ARIMA version.

In [13] Ai et al. The use of guide vector machines that allows you to forecast how often human beings gets certain cancers, the researchers used a aid vector gadget (SVM) version. According to the findings, the SVM version effectively anticipated the incidence of numerous cancers.

In [14] Dong and Taslimitehrani using a sample-aided regression modelling approach. Disease prevalence costs have been analysed in connection to many demographic parameters the use of a regression model in this studies. The results established the regression version's ability to correctly forecast the improvement of numerous illnesses.

In [15] Ali et al. reviewed the cutting-edge kingdom of most cancers in India and offered hints for the destiny of most cancers research. They have a look at addressed the progress of most cancers studies in India and its gift kingdom. The authors proposed some of changes they believe will enhance most cancers take a look at and manipulate efforts in India.

In end, these studies display how diverse approaches, consisting of statistical fashions and machine gaining knowledge of algorithms, can be used to forecast most

ISSN: 2366-1313

cancers prevalence and loss of life costs throughout geographic areas. The results imply a growing most cancers burden in most locations, underscoring the want of efficient most cancers prevention and control efforts. There is promise for combining on-line inquiries and gadget getting to know algorithms to improve the precision of most cancers forecasts, as is proven by way of the studies.

III METHODOLOGY

Dataset

Dataset together with information on most cancers-related deaths international from 1990 to 2019. This records series presents international estimates of most cancers costs, deaths, and different parameters related to the ailment. This records comes from the Worldwide Burden of Disease venture and is supposed for researchers, analysts, and policymakers interested by examining international cancer developments and styles. Multiple documents make up the dataset, which include information on hazard factors, most cancers occurrence, and mortality; years of existence lost, years of existence adjusted for disability, and years lived with disability. The wide variety of those who died from each type of most cancers between 1990 and 2019 is protected within the 02 general-cancer-deaths-by way of-kind. Csv record. The information is separated into classes based on age, gender, and location. Researchers and analysts interested in the worldwide burden of most cancers and in spotting styles and traits in cancer mortality may locate this dataset beneficial.

Link:

https://www.Kaggle.Com/datasets/belayethossainds/most cancers- and-deaths-dataset-19902019-globally

Regression Model

Modelling the connection among a based variable and a hard and fast of unbiased variables is a not unusual use of regression analysis, a popular statistical technique. Its cause is to foretell a continuous numeric cost, which includes a destiny promoting price or quantity of a product. Linear regression, selection tree regression, random woodland regression, and polynomial regression are just a few examples of the numerous varieties of regression fashions" to be had.

Linear regression is the most common and easiest to apply regression approach. It presupposes that the relationship among the explanatory and explanatory variables is linear. That is, if you regulate the unbiased variable, the based variable will shift in kind. The purpose of linear regression is to discover the smallest viable deviation between the expected and discovered values via modelling the

ISSN: 2366-1313

relationship among them the usage of a instantly line.

The use of choice bushes to do regression analysis is called non-parametric choice tree regression. A preference can be reached regarding the structured variable by further subdividing the facts relying on the impartial elements. The choice tree is built by using iteratively breaking the data into subsets in keeping with a predetermined set of criteria designed to increase the homogeneity of each subset. The final product seems like a tree, with every leaf representing a price for the established variable as anticipated by using the version.

As an ensemble technique, random forest regression uses a set of character selection timber to provide extra reliable forecasts. The algorithm works through generating many choice bushes, each of which makes use of an extraordinary subset of the records and the impartial variables at random. After all decision timber has made their forecasts, a median prediction is calculated.

The link between the dependent and independent variables is modelled as a polynomial of nth degree in polynomial regression. It is utilized in situations whilst the connection between the elements isn't linear. Complex statistics styles may be overlooked by using linear regression but no longer through polynomial regression. Over fitting, which can also cause inaccurate predictions when implemented to sparkling facts, is much more likely to arise the usage of this technique.

Evaluation Parameters

It is feasible to gauge a regression version's efficacy with the aid of adjusting its regression parameters. Some of the maximum famous regression parameters and their definitions comply with:

•R2 Score (Coefficient of Determination): The R2 score quantifies the goodness-of-fit between the data and the regression model. It indicates what fraction of the variation in the based variable can be accounted for through the impartial variable (s). A wonderful R2 score suggests that the model fits the information exactly.

€The percentage of variant within the established variable that can be attributed to the unbiased variable is known as its "Explained Variance Score," or "EVS" (s). Also among 0 and 1, with 1 indicating that each one of the variant in the dependent variable is defined by means of the model.

Mean Squared Error (MSE) is a statistical degree of the way off your estimates are from the target values. When the MSE is less, the model appears to be a better in shape for the information.

Mean Absolute Error (MAE) is a

ISSN: 2366-1313

statistical degree of ways some distance off the anticipated value of a variable is from the real price. Comparatively much less liable to outliers than MSE, it's miles a decent indication of the version's correctness.

Root Error in Mean Squared (RMSE) Units are the same as the based variable given that RMSE is the square root of MSE. It is common guidance to evaluate the effectiveness of several regression models by way of calculating their root-mean-squared errors (RMSEs).

IV INDIAN CANCER FORCASTING METHODOLOGY

Each component of the block diagram for predicting the number of deaths from cancer in India using supervised machine learning is described in detail below. Polynomial regression, which is less cumbersome and quicker than standard methods, has also been conducted in this article.

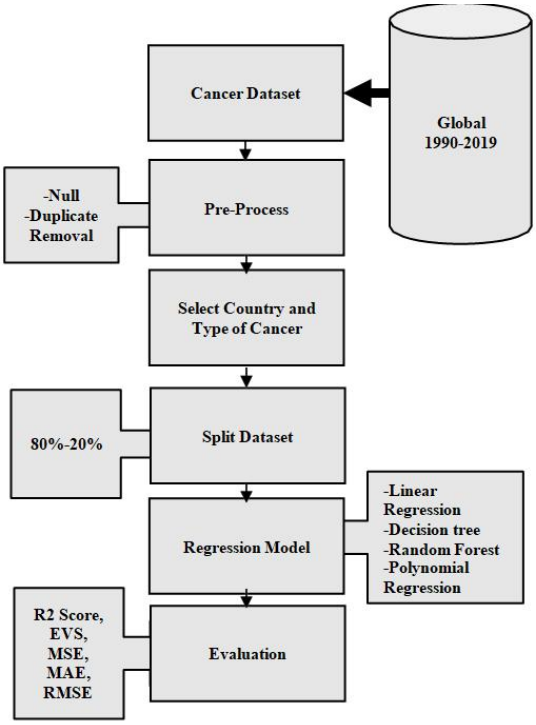

Fig. 1. Lung Sound Reorganization Proposed Block Diagram

The proposed explains how to construct a regression model to forecast cancer incidence rates.

Step 1: Download the cancer dataset available on Kaggle, which spans the globe from 1990 to the 2019 year.

Step 2: Pre-processing the dataset involves deleting duplicate or null values and replacing them with new ones. This is done using pandas library ***dropna()*** and ***drop duplicates()*** Functions.

Step 3: After narrowing down the dataset to a certain nation and cancer kind, the dataset is divided 80:20 between a training set and a testing set.

Step 4: This step will now use Linear Regression, Decision Tree, Random Forest, and Polynomial Regression are the four

ISSN: 2366-1313

types of regression models that are trained using the trainingset.

Step 5: Five independent metrics—the R2 score, the Explained Variance Score (EVS), the Mean Squared Error (MSE), the Mean Absolute Error (MAE), and the Root Mean Squared Error—are then used to assess the quality of the regression models (RMSE). These measures are used to rank the various regression models and decide which one will provide the most accurate forecasts.

V RESULTS AND EVALUATION

Here, we predict cancer mortality in India the usage of a publicly reachable dataset from the Global Burden of Disease Study and three wonderful supervised learning strategies, inclusive of linear regression, choice tree regression, random wooded area regression, and polynomial regression. We used a spread of signs in our version evaluations.

```
import pandas as pd
df=pd.read_csv("/content/02 total-cancer-deaths-by-type.csv")
pd.set_option('display.max_columns', None)
df.head()
```

	Entity	Code	Year	Deaths - Liver cancer - Sex: Both - Age: All Ages (Number)	Deaths - Kidney cancer - Sex: Both - Age: All Ages (Number)	Deaths - Lip and oral cavity cancer - Sex: Both - Age: All Ages (Number)	Deaths - Tracheal, bronchus, and lung cancer - Sex: Both - Age: All Ages (Number)	Deaths - Larynx cancer - Sex: Both - Age: All Ages (Number)	Deaths - Gallbladder and biliary tract cancer - Sex: Both - Age: All Ages (Number)	Deaths - Malignant skin melanoma - Sex: Both - Age: All Ages (Number)	Deaths - Leukemia - Sex: Both - Age: All Ages (Number)	Deaths - Hodgkin lymphoma - Sex: Both - Age: All Ages (Number)	Deat Mult mye - Bo Age: (Num
0	Afghanistan	AFG	1990	851	66	89	983	260	180	47	1055	102	
1	Afghanistan	AFG	1991	866	66	89	982	263	182	48	1089	108	
2	Afghanistan	AFG	1992	890	68	91	989	268	185	51	1171	118	
3	Afghanistan	AFG	1993	914	70	93	995	275	189	53	1252	126	
4	Afghanistan	AFG	1994	933	71	94	996	282	193	54	1296	130	

Fig. 2. Data Reading

```
TYPES_OF_CANCER = df.columns[3:]
print(TYPES_OF_CANCER)

Index(['Deaths - Liver cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Kidney cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Lip and oral cavity cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Tracheal, bronchus, and lung cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Larynx cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Gallbladder and biliary tract cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Malignant skin melanoma - Sex: Both - Age: All Ages (Number)',
       'Deaths - Leukemia - Sex: Both - Age: All Ages (Number)',
       'Deaths - Hodgkin lymphoma - Sex: Both - Age: All Ages (Number)',
       'Deaths - Multiple myeloma - Sex: Both - Age: All Ages (Number)',
       'Deaths - Other neoplasms - Sex: Both - Age: All Ages (Number)',
       'Deaths - Breast cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Prostate cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Thyroid cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Stomach cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Bladder cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Uterine cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Ovarian cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Cervical cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Brain and central nervous system cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Non-Hodgkin lymphoma - Sex: Both - Age: All Ages (Number)',
       'Deaths - Pancreatic cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Esophageal cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Testicular cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Nasopharynx cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Other pharynx cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Colon and rectum cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Non-melanoma skin cancer - Sex: Both - Age: All Ages (Number)',
       'Deaths - Mesothelioma - Sex: Both - Age: All Ages (Number)'],
      dtype='object')
```

Fig. 3. Types of Cancer

```
R2 Score= 0.9539860476933492
EVS= 0.963877340103035
MSE= 1423377423.573189
MAE= 30785.898118279874
RMSE= 37727.67450523805
<matplotlib.legend.Legend at 0x7f428d6dc160>
```

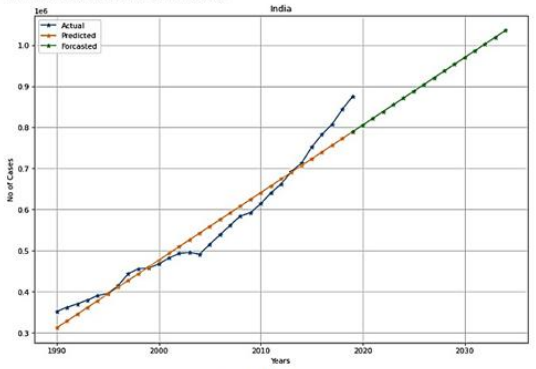

Fig. 5. LR Regression

As shown in figure 6 shows Years in x-axis and no of cancer cases in India at y-axis. In that linear regression model is train using 1990 to 2019 data and after forecast future dataup to 2035 year.

```
R2 Score= 0.9993472143923591
EVS= 0.9993663522471034
MSE= 20193012.11418283
MAE= 3665.468705495198 7
RMSE= 4493.663551511487
<matplotlib.legend.Legend at 0x7f428a236c20>
```

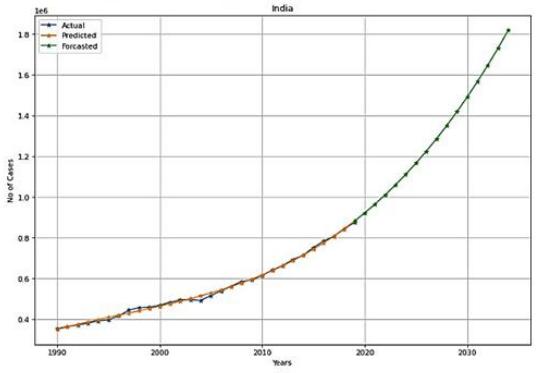

Fig. 8. Polynomial Regression

ISSN: 2366-1313

As shown in figure 8 shows Years in x-axis and no ofcancer cases in India at y-axis. In that polynomial regression model is train using 1990 to 2019 data and after forecasting future data up to 2035 year.

VI Conclusion

These outcomes endorse that projecting cancer mortality in India is viable the use of supervised machine mastering strategies. The findings of this research may additionally tell the efforts of policymakers and healthcare providers within the fight towards most cancers. However, the generalizability of our findings may be affected by the reality that we handiest used a single dataset and a small subset of attributes. Using a publicly reachable dataset from Kaggle, we implemented 4 wonderful supervised gaining knowledge of algorithms to the hassle of cancer demise prediction in India. These had been linear regression (LR), choice tree regression (DT), random forest regression (RF), and polynomial. The polynomial regression model outperforms the options. The polynomial regression version has an R2 Score, EVS, MSE, MAE, and RMSE of zero.Ninety three, zero.Ninety three, 0.Ninety three, and 0.93, respectively.

In order to decorate the precision of most cancers forecasting in India, future examine would possibly check out the use of extra varied datasets and complicated gadget learning algorithms. The accuracy and application of the forecasting version may be advanced by encompass other parameters, which includes demographic, socioeconomic, and environmental components. Cancer prevention and management in India and different countries with similar problems might also benefit from the software of machine gaining knowledge of in most cancers forecasting.

REFERENCES

1. K. Sathishkumar, M. Chaturvedi, P. Das, S. Stephen, and P. Mathur, “Cancer incidence estimates for 2022 & projection for 2025: Result from National Cancer Registry Programme, India.,” The Indian journal of medical research, vol. 156, no. 4&5, pp. 598–607, 2022, doi: 10.4103/ijmr.ijmr_1821_22.

2. C. Tudor, “A Novel Approach to Modelling and Forecasting Cancer Incidence and Mortality Rates through Web Queries and Automated Forecasting Algorithms: Evidence from Romania.,” Biology, vol. 11, no. 6, Jun. 2022, doi: 10.3390/biology11060857.

3. R. Gupta, A. Sharma, V. Anand, and S. Gupta, “Automobile Price Prediction using

ISSN: 2366-1313

Regression Models," in 2022 International Conference on Inventive Computation Technologies (ICICT), 2022, pp. 410–416. doi: 10.1109/ICICT54344.2022.9850657.

4. M. Dalmartello et al., "European cancer mortality predictions for the year 2022 with focus on ovarian cancer," Annals of Oncology, vol. 33, no. 3, pp.

5. B. Trächsel, E. Rapiti, A. Feller, V. Rousson, I. Locatelli, and J.-L. Bulliard, "Predicting the burden of cancer in Switzerland up to 2025," PLOS Global Public Health, vol. 2, no. 10, p. e0001112, Oct. 2022, [Online]. Available: https://doi.org/10.1371/journal.pgph.0001112

6. K. W. Jung, Y. J. Won, M. J. Kang, H. J. Kong, J. S. Im, and H. G. Seo, "Prediction of Cancer Incidence and Mortality in Korea, 2022," Cancer Research and Treatment, vol. 54, no. 2, pp. 345–351, 2022, doi: 10.4143/crt.2022.179.

7. B. Sekeroglu and K. Tuncal, "Prediction of cancer incidence rates for the European continent using machine learning models," Health Informatics Journal, vol. 27, no. 1, p. 1460458220983878, Jan. 2021, doi: 10.1177/1460458220983878.

8. Prasadu Peddi (2015) "A review of the academic achievement of students utilising large-scale data analysis", ISSN: 2057-5688, Vol 7, Issue 1, pp: 28-35

9. Prasadu Peddi (2015) "A machine learning method intended to predict a student's academic achievement", ISSN: 2366-1313, Vol 1, issue 2, pp:23-37.