Prediction of cancer incidence rates for the European continent using machine learning models

Boran Sekeroglu and Kubra Tuncal
Information Systems Engineering, Near East University, Cyprus

Abstract
Cancer is one of the most important and common public health problems on Earth that can occur in many different types. Treatments and precautions are aimed at minimizing the deaths caused by cancer; however, incidence rates continue to rise. Thus, it is important to analyze and estimate incidence rates to support the determination of more effective precautions. In this research, 2018 Cancer Datasheet of World Health Organization (WHO), is used and all countries on the European Continent are considered to analyze and predict the incidence rates until 2020, for Lung cancer, Breast cancer, Colorectal cancer, Prostate cancer and All types of cancer, which have highest incidence and mortality rates. Each cancer type is trained by six machine learning models namely, Linear Regression, Support Vector Regression, Decision Tree, Long-Short Term Memory neural network, Backpropagation neural network, and Radial Basis Function neural network according to gender types separately. Linear regression and support vector regression outperformed the other models with the $R^2$ scores 0.99 and 0.98, respectively, in initial experiments, and then used for prediction of incidence rates of the considered cancer types. The ML models estimated that the maximum rise of incidence rates would be in colorectal cancer for females by 6%.

Keywords
cancer incidence rates, machine learning, Europe, linear regression, support vector regression

Introduction
The leading cause of abnormal death worldwide is cancer; thus, it is a highly prevalent and significant health problem.1 Researches have been performed for the early detection and treatment of cancer, mortality rates remain extremely high when compared to other fatal diseases. These rates are obtained by the local registration systems of nations and the World Health Organization for different countries and periods.2

Corresponding author:
Boran Sekeroglu, Information Systems Engineering, Near East University, Mersin 10 Turkey, Nicosia, 1403, Cyprus.
Email: boran.sekeroglu@neu.edu.tr

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
Even though registration systems for different types of cancer diseases are implemented in different developed countries, the same cannot be assumed for undeveloped countries. Thus, registered rates do not provide exact statistical information, but provide data to form a general opinion about the incidence and mortality rates of different types of cancer all over the world. In 2018, the WHO published a report that provided statistical data on cancer diseases up to 2012. They identified that the total global population was 7,632,819,272 and that 18,078,957 new cancer cases and 9,555,027 deaths had occurred all over the world from 35 different types of cancer.

The highest incidence rates were recorded for Lung, Breast, Colorectal, Prostate, Stomach and Other Cancer types when both genders were considered for all age groups. In the same report, it was identified that the cancer types with the highest mortality rates were Lung Cancer, Colorectal Cancer, Stomach Cancer, Liver Cancer, Breast Cancer, Esophageal Cancer, Pancreatic Cancer, Prostate Cancer and Other cancers. The same report noted that Europe, as a continent, consisted of 22 countries with registration records and a total population of 743,837,100. It was also identified that the number of new cancer cases was 4,229,662 and the number of resultant deaths was 1,943,478 in 2012 across Europe.

The report showed that the incidence rates of Prostate, Lung, Colorectal, Bladder and Kidney Cancers were higher than other cancer types for men and the incidence rates of Breast, Colorectal, Lung, Corpus Uteri and Melanoma were higher than other types for women within all age groups.

To analyze the mortality rates, the initial step involve the analysis of incidence rates of different cancer types. Researchers have performed different studies to predict the incidence and/or mortality rates to overcome the inconsistency of some data and resolve the problem of missing and inadequate registration records. However, making such predictions is a challenging task. Therefore, researchers tend to use Machine Learning (ML) models for the robust prediction of both incidence and mortality rates.

Different cancer types have different incidence rates in various countries and this makes public awareness and precautions research inefficient. The prediction of incidence rates within different countries could provide valuable information in terms of prevalence and public awareness. This would lead to cost savings for financial and human resources that require significant funding. Additionally, prediction results provide information to researchers about the course of events and the effectiveness of precautions taken for particular cancer types. However, it is not easy for such predictions to be performed by human beings than machine learning models because of their execution times, capacity to make connections between data, and uninterrupted operational ability. Therefore, ML models were implemented almost in every field of our lives to help human experts solve problems or make decisions. In addition, ML and has been used and explored in different fields of health science such as bioinformatics and particularly in cancer predictions to support cancer researchers and the general public for further precautions and awareness.

Senturk and Senturk conducted research on a database obtained from the UCI Machine Learning Repository using Backpropagation Neural Networks (BPNN) and achieved 77% of accuracy rate in breast cancer classification.

Kourou et al. investigated several models to determine the efficiency of machine learning techniques in cancer prognosis and prediction. They concluded that research has mainly focused on supervised models for the development of predictive algorithms. These investigations suggested that analysis could be faster and more efficient estimations could be made with larger data. They also aimed to obtain the best results by using machine learning techniques by excluding human factors.

Mohammadzadeh et al. used decision trees to predict the mortality rate of gastric cancer patients. They used data from 216 patients and achieved a prediction accuracy of 74%. O’Lorcain et al. implemented Log and log-linear Poisson regression models for the prediction of colorectal
cancer. They fit the model using data from the World Health Organization for the period from 1950 to 2002 to predict the mortality rates in Ireland.

Malvezzi et al.\textsuperscript{16} used a Linear Regression model to predict cancer mortality rates across the European Union and six other European countries. Ribes et al.\textsuperscript{17} used Bayesian models to predict both incidence and mortality rates in Catalonia up to 2020. They obtained the data from cancer registries in Spain and Catalonia. Alhaj and Maghari\textsuperscript{18} considered Random Forest and Rule Induction Algorithms to predict the cancer survivability rates in the Gaza Strip. They concluded that Random Forest achieved more accurate results of 74.6% compared with rule induction algorithms.

Recently, Jung et al.\textsuperscript{19} implemented a Jointpoint regression model to predict cancer incidence and mortality rates in Korea for 2019. They used the Korea National Cancer Incidence Database in their research. Malvezzi et al.\textsuperscript{20} undertook comprehensive research on prediction studies. Jointpoint regression was implemented to predict lung cancer rates in Europe.

Although these studies have obtained high prediction or classification results, they only considered several cancer types for specific regions and a limited number of ML models. In this research, six ML models are considered to obtain prediction results for the incidence rates of Lung Cancer, Breast Cancer, Prostate Cancer, Colorectal Cancer and All Other Types of Cancers for 22 European countries using the latest dataset published by the WHO. ML models with different training and testing ratios were utilized. Superior models determined by initial experiments were trained to predict the incidence rates of these five types of cancer for 8 years until 2020 to enable researchers to make forward steps and to create a resource for planning cancer-control and social awareness programs for the future. Therefore, the main objective of this research is to examine ML prediction results for European cancer incidence rates and draw links to how society can benefit from more ML insights to form data in creating effective precautions and to distribute funding resources efficiently.

The rest of the paper is organized as follows. Section 2 introduces the materials and methods used in this research and Section 3 presents the obtained results. Discussions will be presented in Section 4 and finally Section 5 concludes the work done in this research.

Materials and methods

This section briefly introduces the basic characteristics of dataset, the considered machine learning models and presents the details of the design of experiments.

Dataset

The data of the European continent were obtained from the World Health Organization 2018 Data sheet\textsuperscript{3} (http://ci5.iarc.fr) It includes incidence rates of 29 cancer types for 22 European countries including Austria (with 3 regions), Bulgaria, Belarus, Croatia, Cyprus, Denmark, Estonia, France (with 9 regions), Germany (with 2 regions), Iceland, Ireland, Italy (with 8 regions), Lithuania, Malta, Netherlands, Norway, Poland, Slovakia, Slovenia, Spain (with 9 regions), Switzerland (with 6 regions) and the UK (with 11 regions). Male and female groups are considered separately.

The start dates of the records vary from 1953 to 1998 and most of the records are ended in 2012. The records of only two countries, Italy and Slovakia, were ended in 2010. However, since the majority of the countries started to register cancer incidence records from the beginning of the 1990s, 1993 was selected as the start date of incidence records to minimize the impact of a significant amount of missing values that could affect the accuracy of the prediction results. Therefore, data for 19 years for 22 countries were used, with a total of 418 input data for each cancer type in each experiment.
The number of missing values was reduced by replacing them with the nearest neighbor data imputation technique. Missing years for Italy and Slovakia (2011 and 2012) were replaced by values from 2010. Data normalization was performed by a Minimum-Maximum scaler to normalize data between 0 and 1 for each attribute.

Four cancer types with the highest incidence rates for male and female groups were considered in this research, namely Lung Cancer, Breast Cancer, Prostate Cancer and Colorectal Cancer. In addition to these, the total incidence rates of the remaining 22 cancer types were considered for All Cancer Types.

For countries that had different record regions, to reduce the complexity of the machine learning models, the region that had the maximum incidence rate was selected to represent the whole country.

Evaluation of the obtained results was performed according to two criteria to determine the optimum ML model for future prediction. These criteria are Mean Squared Error (MSE) and $R^2$ Score which are the main indicators of the success of prediction results of models.21

**Machine learning models**

Several prediction models are possible through machine learning. Some of them are directly related to basic or advanced statistics, while others are associated with the field of neural networks, which tries to simulate human perception, or tree based algorithms etc. Each model has unique principles with several advantages and disadvantages. If there is a non-linear relationship between the data attributes, the performance of statistical methods may be reduced, and this might cause ineffective predictions. However, for linear data types, they can provide optimal results. Neural network models defined as black-boxes because of their internal structures and hidden layers. However, most researchers have concluded that if the number of instances of observed data is huge and non-linear relationships exist between attributes, neural network models can produce superior results compared to other ML models. On the other hand, when there is not a significant number of instances, they fail to produce efficient or accurate results.

For this reason, several comparative studies have been performed to determine superior ML models for prediction problems.22–24 However, researchers have concluded that there is no single model that is superior for all kinds of prediction or classification problems and also different researchers have determined different ML models to be superior in different experiments.

Hashem et al.22 conducted a research to predict liver fibrosis in chronic hepatitis C patients. They considered particle swarm optimization, decision tree, multi-linear regression and genetic algorithm models in their research. They concluded that the alternating decision tree model was the superior model for the considered task. Mao et al.23 performed experiments to predict human comfort. They employed support vector regression with a radial basis function kernel, and compared the obtained results with linear regression, forms of ridge regression, support vector regression with linear kernel and multi-layer perceptron. They concluded that SVR with Radial Basis Function kernel achieved the highest prediction results. Kirsal Ever et al.24 performed large-scale comparative study on four ML models, namely back-propagation neural network, radial-basis function neural network, support vector regression (SVR) with radial basis function kernel and decision tree, for nine different prediction problems. They concluded that the neural network algorithms, achieved higher rates and consistent results than other considered models.

When all these researches are considered, it is evident that researchers tend to implement more than one model for their research needs and generally each research can involve a comparative study between different models.
In this research, we used the six frequently applied models to achieve optimal results, namely, linear regression, decision tree, support vector regression, and three neural network models, long-short term memory, backpropagation and radial-basis function neural networks.

Linear Regression is a basic statistical method that draws a best-fitting regression line through real and predicted points. It is frequently used in prediction problems, especially on datasets that have linear correlation between attributes.

Decision Trees are another supervised and effective kind of algorithm used for both classification and prediction. They form a hierarchical relationship between instances and attributes starting from a root node to decision nodes. A sequence of questions is presented in the form of a tree structure. Leaf nodes indicate the classification or prediction results.

Support Vector Regression is a special kind of Support Vector Machines. They are designed to accept real-valued numbers and to perform prediction instead of classification. It is widely and effectively used for prediction problems.

LSTM is a type of recurrent neural network in which the internal structure allows it to remember previous knowledge that is considered for future predictions. It is used extensively for prediction problems.

Backpropagation neural networks are some of the most basic and widely used neural networks for both prediction and classification problems. They form the basis for more advanced researches such as deep learning.

Radial-basis Function neural networks are similar to Backpropagation but use a radial-basis function in a constant single hidden layer to provide effective and efficient results for both classification and prediction researches.

Design of experiments

Different training ratios can produce various prediction performances; therefore, the six machine learning models described above were trained by using 60%, 70%, and 80% training ratios to determine the most efficient ratio for future prediction.

During the training of the models, experiments were divided into two groups: Male and Female. Then, all the obtained data were analyzed according to the evaluation criteria, namely the Mean Squared Error (MSE) and $R^2$ Score of each group.

Each machine learning model has its own unique hyper-parameter to increase learning ability and prediction performance. For decision tree, attribute selection criterion was used as the mean squared error which is used for prediction problems. In support vector regression, the most frequently used kernel function, radial basis function was used. After several experiments, parameter values $\gamma$ and $\varepsilon$ that affect the prediction accuracy of SVR were determined as 0.005 and 0.01 respectively.

In the backpropagation neural network, two hidden layers were used with the Sigmoid Activation Function for each of the following experiments. Optimum results were obtained by 500 hidden units in each hidden layer. Maximum iterations were limited to 3000 in order to avoid over-fitting which provides perfect convergence of training data but ineffective predictions on testing data.

In the radial basis function neural network, the learning rate was determined as 0.09 and maximum iterations were limited to 4000 in order to avoid over-fitting.

In LSTM, three hidden layers were added to the architecture to increase the prediction ability of the model. In the output layer, Sigmoid Activation Function was used and maximum iterations was limited to 200.

After determining the optimum models for each cancer type, future prediction was performed by re-training the dataset with the most effective training ratio with estimation performed for the
(untrained) next year. The estimated year was added to the dataset and this process was repeated until 2020. Figure 1 demonstrates the block diagram of the future estimation process.

**Experimental results**

Experimental results are divided into two categories: model evaluation experiments and future estimation results. Model evaluation experiments were performed to analyze the prediction performances of models and then, optimum models were used for future prediction of cancer incidence rates.

The incidence rates that were recorded by registration systems, indirectly include the annual information for population, factors that affect rates, the success of the precautions taken etc. In time-series prediction, all these factors are assumed to have similar distributions in terms of years and this makes it possible to create a relation to future years by machine learning models except for unexpected extreme factors. Therefore, the incidence rate of each corresponding year provides a linear or non-linear relation between the previous and following years. The success of the prediction results was tested on the untrained data (test set), which were the last 5 years for 80% of the training ratio, the last 7 years for 70% of the training ratio and the last 9 years for 60% of the training ratio. The predicted results were compared with actual incidence rates in these test sets and the results were evaluated by two performance evaluation criteria mentioned in the text. While the predicted results of the test years demonstrated the efficiency of optimum models for the prediction of untrained data, estimated years were added to the dataset to fill out the missing years until 2020.

**Model evaluation experiments**

As indicated above, six models were trained for the prediction of cancer incidences for five different types of cancer with different gender groups for the European countries. In the following subsections, the results will be presented.

The cancer types considered for the Female group were Lung, Breast, Colorectal and All Cancer Types. In this case, Linear Regression Model outperformed other models in all experiments when $R^2$ and $MSE$ results were considered. Table 1 shows the optimum and closest results of the Female Group experiments and indicates the training ratios used to obtain optimum results. Figure 2 shows the prediction graph of the Linear Regression Model for Breast Cancer using 60% of the training ratio on the test (untrained) years.

In the Male Group experiments, different results were obtained from the Female Group results. Linear Regression and Support Vector Regression produced superior results in different cancer types. Linear Regression outperformed other models in Lung Cancer and All Cancer Types experiments, however in the rest of the experiments, Support Vector Regression produced more accurate
Table 1. Optimum results of female experiments.

| Cancer Type   | Model         | Training ratio % | $R^2$ Score | MSE        |
|---------------|---------------|------------------|-------------|------------|
| Lung cancer   | Linear regression | 70               | 0.9834      | 9.95x10^{-5} |
|               | RBFNN         | 70               | 0.9600      | 3.7x10^{-3}  |
| Breast cancer | Linear regression | 60               | 0.9892      | 5.14x10^{-5} |
|               | SVR           | 60               | 0.9670      | 2.8x10^{-3}  |
| Colorectum cancer | Linear regression | 80               | 0.9814      | 9.25x10^{-4} |
|               | SVR           | 70               | 0.9810      | 2.3x10^{-3}  |
| All cancer types | Linear regression | 60               | 0.9987      | 1.52x10^{-5} |
|               | SVR           | 60               | 0.9940      | 9.0x10^{-5}  |

Figure 2. Prediction graph of linear regression model for breast cancer with 60% of training ratio.

Table 2 shows the optimum and closest results of the Male Group experiments and indicates the training ratios that obtained optimum results.

Figures 3 and 4 show the Prediction Graphs of the Linear Regression and Support Vector Regression for different experiments of the Male Group on test (untrained) years respectively.
Future prediction results for lung cancer

Estimation was performed with the Linear Regression Model for both genders. In the Female Group, only Cyprus had increasing incidence rates for Lung Cancer in 2020 with 5.0%. Although

| Model                  | Training ratio % | $R^2$ Score | MSE    |
|------------------------|------------------|-------------|--------|
| Optimum result         | Linear regression| 70          | 0.9976 | $3.79 \times 10^{-5}$ |
| Followed by            | SVR              | 80          | 0.9890 | $1.3 \times 10^{-3}$  |

| Model                  | Training ratio % | $R^2$ Score | MSE    |
|------------------------|------------------|-------------|--------|
| Optimum result         | SVR              | 80          | 0.9891 | $1.2 \times 10^{-3}$  |
| Followed by            | RBFNN            | 80          | 0.9520 | $4.7 \times 10^{-3}$  |

| Model                  | Training ratio % | $R^2$ Score | MSE    |
|------------------------|------------------|-------------|--------|
| Optimum result         | SVR              | 80          | 0.9890 | $1.2 \times 10^{-3}$  |
| Followed by            | RBFNN            | 70          | 0.9120 | $7.6 \times 10^{-3}$  |

| Model                  | Training ratio % | $R^2$ Score | MSE    |
|------------------------|------------------|-------------|--------|
| Optimum result         | Linear regression| 80          | 0.9975 | $3.42 \times 10^{-5}$ |
| Followed by            | SVR              | 70          | 0.9911 | $1.0 \times 10^{-4}$  |

Figure 3. Prediction graph of linear regression model for lung cancer (male) with 70% of training ratio.

Future prediction results for lung cancer

Estimation was performed with the Linear Regression Model for both genders. In the Female Group, only Cyprus had increasing incidence rates for Lung Cancer in 2020 with 5.0%. Although
all other European countries had decreasing incidence rates, the rates significantly increased between 2015 and 2018, but then the rates decreased below 2012 limits.

In the Male Group, most of the European countries had increasing incidence rates for Lung Cancer. Only Austria, Bulgaria, Estonia, Italy and Switzerland had decreased incidence rates. Table 3 shows the countries with increasing incidence rates and the corresponding percentages for the Male group.

**Future prediction results for breast cancer**

This estimation was performed using the Linear Regression Model. When the incidence rates of 2012 were considered, only two countries, Switzerland and Italy had increasing incidence rates for Breast Cancer in 2020 of 1.0% and 1.8% respectively. However, fluctuations occurred for Austria during the period between 2014 and 2019. While it has a reduction in incidence rates in 2020, an increase starts after 2018. All other countries have stable reductions of incidence rates during this period.

**Future prediction results for prostate cancer**

This estimation was performed using the Support Vector Regression Model. A general reduction is predicted in all European countries for Prostate Cancer with two exceptions, namely Malta and

---

**Table 3.** Countries and percentages of raising incidence rates of lung cancer in 2020 (male group).

| Country   | Increase rate % | Country | Increase rate % | Country | Increase rate % |
|-----------|-----------------|---------|-----------------|---------|-----------------|
| Belarus   | 5.0             | Malta   | 9.2             | Croatia | 2.5             |
| Netherlands | 4.8           | Cyprus  | 5.0             | Norway  | 1.4             |
| Denmark  | 0.6             | Poland  | 4.1             | France  | 3.7             |
| Slovakia | 1.2             | Germany | 2.8             | Slovenia | 1.0             |
| Iceland  | 15.5            | Spain   | 3.1             | Ireland | 1.2             |
| UK       | 1.8             | Lithuania | 3.9          | –       | –               |
Spain, which exhibit increases of 7.5% and 3.0% respectively. A decrement is also predicted for Austria with unstable incidence rates predicted between 2014-2018.

**Future prediction results for colorectum cancer**

This estimation was performed using the Linear Regression Model and the Support Vector Regression Model for Female and Male Groups respectively.

A significant increase was predicted in all European countries for the Colorectal cancer in Female Group. Almost all countries had increasing incidence rates, except for Austria, Denmark and Iceland. Table 4 shows the countries with increasing incidence rates and the corresponding percentages for the Female group for Colorectal Cancer.

Contrary to the Female Group estimation results, decreased incidence rates were estimated for the Male Group for Colorectal Cancer. Only one country, Austria, had increasing incidence rates with a rate of increase of 1.5%.

**Future prediction results for all cancer types**

This estimation was performed using the Linear Regression Model for both genders. It was estimated that most of the European countries would have a decrease in incidence rates for all cancer types in 2020 for the Female Group. However, Austria, Bulgaria, Belarus, Denmark, Germany, Ireland, Slovenia and Spain will have minor increase rates in 2020. In the Male Group, it was estimated that Austria, Bulgaria, Estonia, France, Italy and Lithuania would have increasing incidence rates for all cancer types in 2020.

However, the change in incidence rates in terms of both increment and decrement, will not exceed 1.0% for either gender in all European countries.

Figures 5 and 6 presents the prediction results for Cyprus (female) and the UK (men) for all considered cancer types respectively. The vertical line in the graphs represents the boundary for the year when the estimation starts. Table 5 presents the predicted total incidence rates (number of persons) for European countries for males and females separately.

**Discussions**

The obtained results should be analyzed in two ways: the analysis of machine learning models and the prediction of incidence rates for both genders respectively.
The results show that unstable results occurred in the Female Group experiments except for the Linear Regression Model. A more linear relationship was observed in the Female data between the years and the cancer incidence rates; therefore, the Linear Regression model produced optimum results for all experiments in this group. SVR produced more accurate results in terms of $R^2$ Scores for some experiments; however, BP and RBFNN produced close, better or equal results compared to SVR. Decision Tree minimized errors while increasing training samples, but this did not improve the prediction ability of the model.

**Figure 5.** Future estimation graph of Cyprus (female) for all cancer types.

**Figure 6.** Future estimation graph of UK (male) for all cancer types.

The results show that unstable results occurred in the Female Group experiments except for the Linear Regression Model. A more linear relationship was observed in the Female data between the years and the cancer incidence rates; therefore, the Linear Regression model produced optimum results for all experiments in this group. SVR produced more accurate results in terms of $R^2$ Scores for some experiments; however, BP and RBFNN produced close, better or equal results compared to SVR. Decision Tree minimized errors while increasing training samples, but this did not improve the prediction ability of the model.
In the male group experiments, non-linear increment and decrement of rates in some cancer types and countries across the years led to the failure of the Linear Regression model to produce optimum results in some experiments. In these experiments, SVR, LR, and neural-based model RBFNN were found to be superior; however, RBFNN could not produce optimal results in any of the experiments. Similar to the Female Group results, DT minimized the errors but could not provide effective predictions. Although LSTM is considered one of the superior models for prediction tasks, it was not able to make efficient predictions in either the Male and Female Experiments. Generally, it needs large volume of data to minimize errors and to remember previous experiences during the training phase, and to perform efficient predictions.

Another important observation is the training ratios for the produced optimum results for the superior models. In the Female Group, relatively less training ratios were required to produce optimum results. However in the Male group, it was observed that increments in training ratios produced higher prediction results.

After considering the minimum $MSE$ and maximum $R^2$ Scores, superior models performed predictions for the years 2013–2020. In the Female Group Predictions, decreases were estimated for Lung and Breast Cancer in all European countries. Similar to these results, the incidence rates of total cancer cases were expected to decrease in the Female Group. However, dangerous and serious increases were estimated in the colorectal cancer. When Europe as a whole is considered for the Female Group, ML models estimated that there would be a 6.01% total increment in colorectal cancer. However, total decreases of 0.76%, 2.66% and 0.01% were estimated for Lung Cancer, Breast Cancer and the total cancer cases respectively.

In the Male group, an increase in incidence rates was only estimated for Lung cancer. For Prostate, Colorectal and All Cancer Types there was a general decrement in the Male Group. Considering the results of all European countries for the Male Group, it was identified that there would be a 1% increment in Lung Cancer and 5.48%, 4.53%, 0.26% decrements in Prostate Cancer, Colorectal Cancer and total cancer cases respectively.

**Conclusion**

The estimation of cancer incidence rates is vital while it directly associated with mortality rates. In this research, six machine learning algorithms were trained to determine the most effective models to estimate the incidence rates of lung, breast, prostate, colorectal, and total of all cancer types, which have the highest incidence and mortality rates of all 29 cancer types.

Model Evaluation Experiments showed that a single model was not sufficient to estimate all kind of cancer types and data. Therefore, superior models for each cancer type and gender were
used for the estimation. Estimation was performed from 2013 to 2020 by performing estimation year by year and the training the superior models again.

Estimated incidence rates showed that, there will be a rise in incidence rates in most European countries in Colorectal cancer for Female while a reduction is estimated for other cancer types. For Males, the incidence rates of the majority of European countries in Lung cancer will rise in 2020 but there will be decrease in other types.

Because of the effects of the treatments and precautions, some decreased and increased rates balanced each other thus, when all cancer types were considered, similar incidence rates were estimated. However, it should be noted that more than 420,000 persons were predicted to have cancer in 2020 in the considered countries and regions of Europe.

These results could be used for increasing awareness of some cancer types and cancer prevention measures especially for Breast Cancer and Lung Cancer for Female and, Prostate Cancer and Colorectal Cancer for Male.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Boran Sekeroglu https://orcid.org/0000-0001-7284-1173

References
1. Haenssle H, Fink C, Schneiderbauer R, et al. Reader study level-i and level-ii groups, man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *J Ann Oncol* 2018; 29(8): 1836–1842.
2. WHO. *World health organization report*. Geneva: World Health Organization, 2018.
3. Bray F, Colombet M, Mery L, et al. Cancer incidence in five continents, vol. xi (electronic version), http://ci5.iarc.fr (2018, accessed November 2019).
4. Oytun M, Tinazci C, Sekeroglu B, et al. Performance prediction and evaluation in female handball players using machine learning models. *IEEE Access* 2020; 8: 116321–116335.
5. Sekeroglu B and Ozsahin I. Detection of covid-19 from chest x-ray images using convolutional neural networks. *SLAS Technol* 2020; 25(6): 553–565.
6. Alpan K and Igi GS. Classification of diabetes dataset with data mining techniques by using weka approach. In: *4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Istanbul, Turkey, 22–24 October 2020, pp.1–7. New York, NY: IEEE.
7. Zhu S, Sun J, Liu Y, et al. The air quality index trend forecasting based on improved error correction model and data preprocessing for 17 port cities in china. *Chemosphere* 2020; 252: 126474.
8. Brismi T, Xu T, Wang T, et al. Predicting diabetes-related hospitalizations based on electronic health records. *Stat Methods Med Res* 2019; 28(12): 3667–3682.
9. Senanayake S, White N, Graves N, et al. Machine learning in predicting graft failure following kidney transplantation: a systematic review of published predictive modelss. *Int J Med Inform* 2019; 130: 103957.
10. Basu S and Narayanaswamy RA. Prediction model for uncontrolled type 2 diabetes mellitus incorporating area-level social determinants of health. *Med Care* 2019; 57(8): 592–600.
11. Venkatesh R, Balasubramanian C and Kahappan M. Development of big data predictive analytics model for disease prediction using machine learning technique. *J Med Syst* 2019; 43(8): 272.
12. Senturk Z and Senturk A. Breast cancer prediction using neural networks. *El-Cezeri J Sci Eng* 2016; 3(2): 345–350.
13. Kourou K, Exarchos T and Exarchos K. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015; 13: 8–17.
14. Mohammadzadeh F, Noorkojuri H and Pourhoseingholi MA. Predicting the probability of mortality of gastric cancer patients using decision tree. *Irish J Med Science* 2015; 184: 277–284.
15. O’Lorcain P, Deady S and Comber H. Mortality predictions for colon and anorectal cancer for Ireland 2003–17. *Colorectal Dis* 2006; 8: 393–401.
16. Malvezzi M, Bertuccio P, Levi F, et al. European cancer mortality predictions for the year 2014. *Ann Oncol* 2014; 25: 1650–1656.
17. Ribes J, Esteban L, Cleries R, et al. Cancer incidence and mortality projections up to 2020 in Catalonia by means of bayesian models. *Clin Transl Oncol* 2014; 16: 714–724.
18. Alhaj MAM and Maghari AYA. Cancer survivability prediction using random forest and rule induction algorithms. In: *Proceedings of 8th international conference on information technology (ICIT)*, Amman, Jordan, 17–18 May 2017.
19. Jung KW, Won YJ, Kong HJ, et al. Prediction of cancer incidence and mortality in Korea. *Cancer Res Treat* 2019; 51(2): 431–437.
20. Malvezzi M, Bosetti C and Rosso T. Lung cancer mortality in European men: trends and predictions. *Lung Cancer* 2013; 80: 138–145.
21. Sekeroglu B, Dimililer K and Tuncal K. Student performance prediction and classification using machine learning algorithms. In: *8th International conference on educational and information technology*, Cambridge, UK, 2–4 March 2019. ICEIT.
22. Hashem D, Esmat G, Elakel W, et al. Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients. *IEEE/ACM Trans Comput Biol Bioinform* 2018; 15(3): 861–868.
23. Mao F, Zhou X and Song Y. Environmental and human data-driven model based on machine learning for prediction of human comfort. *IEEE Access* 2019; 7: 132910–132922.
24. Kirsal Ever Y, Dimililer K and Sekeroglu B. Comparison of machine learning techniques for prediction problems. In: *Workshops of the international conference on advanced information networking and applications*, 2019, pp. 713–723. Springer.
25. Freedman DA. *Statistical models: theory and practice*. Cambridge: Cambridge University Press, 2009.
26. Araiza-Aguilar JA, Rojas-Valencia MN and Aguilar-Vera RA. Forecast generation model of municipal solid waste using multiple linear regression. *Glob J Env Sci Manag* 2019; 6(1): 1–14.
27. Paul RK, Vennila S, Bhat MN, et al. Prediction of early blight severity in tomato (Solanum lycopersicum) by machine learning technique. *Indian J Agr Sci* 2019; 89(11): 169–175.
28. Chen KM and Chang CC. Error-free separable reversible data hiding in encrypted images using linear regression and prediction error map. *Multimedia Tools Apps* 2019; 78(22): 31441–31465.
29. Dougherty G. *Pattern recognition and classification*. New York, NY: Springer, 2013.
30. Cortes C and Vapnik V. Support-vector networks. *Mach Learn* 1995; 20: 273–297.
31. Hearst MA. Support vector machines. *IEEE Intell Syst* 1998; 13(4): 18–28.
32. Henrique BM, Sobreiro VA and Kimura H. Stock price prediction using support vector regression on daily and up to the minute prices. *J Finance Data Sci* 2018; 4(3): 183–201.
33. Azeez OS, Pradhan B and Shafriz HZM. Vehicular co emission prediction using support vector regression model and GIS. *Sustainability* 2018; 10(10): 3434.
34. Yadav V and Nath S. Daily prediction of pm10 using radial basis function and generalized regression neural network. In: *2018 Recent Advances on Engineering, Technology and Computational Sciences*, Allahabad, RAETCS, 6–8 February 2018.
35. Ping L, Jin W, Sangaiah AK, et al. Analysis and prediction of water quality using lstm deep neural networks in iot environment. *Sustainability* 2019; 11(7): 2058.
36. Wang Y. A new concept using lstm neural networks for dynamic system identification. In: *2017 American Control Conference (ACC)*, Seattle, WA, 24–26 May 2017. New York, NY: IEEE.
37. Qing F and Liu C. Forecasting single disease cost of cataract based on multivariable regression analysis and backpropagation neural network. *Inquiry* 2019; 56: 0046958019880740.

38. Elgin Christo VR, Nehemiah HK, Minu B, et al. Correlation-based ensemble feature selection using bioinspired algorithms and classification using backpropagation neural network. *Comput Math Methods Med* 2019; 2019: 1–17.

39. Pan NK and Lv C. Forecasting primary energy requirements of territories by autoregressive integrated moving average and backpropagation neural network models. *Math Probl Eng* 2019; 2019: 1–14.

40. Dai S, Li L and Li Z. Modeling vehicle interactions via modified lstm models for trajectory prediction. *IEEE Access* 2019; 7: 38287–38296.

41. Rout M, Majhi B, Mohapatra UM, et al. Stock indices prediction using radial basis function neural network. In: *International conference on swarm, evolutionary, and memetic computing*, 2012, pp.285–293.