Lung Cancer Incidence Prediction Using Machine Learning Algorithms

Kubra Tuncal, Boran Sekeroglu, and Cagri Ozkan
Near East University, Information Systems Engineering, Nicosia, TRNC, Mersin 10, Turkey
Email: {kubra.tuncal, boran.sekeroglu, cagri.ozkan}@neu.edu.tr

***Abstract*—Everyday, the frequency of incidence of cancer disease is rising. It is one of the most fatal diseases in the world with several types and there is a few reliable data about incidence and mortality rates of cancer and its types. Thus, the prediction of the rates becomes challenging task for human beings. For this reason, several machine learning algorithms have been proposed to provide effective and rapid prediction of uncertain raw data with minimized error. In this paper, Support Vector Regression, Backpropagation Learning Algorithm and Long-Short Term Memory Network is used to perform lung cancer incidence prediction for ten European countries those records have been started from 1970. Results show that the prediction of incidence rates is possible with high scores with all algorithms; however, Support Vector Regression performed superior results than other considered algorithms.**

***Index Terms*—lung cancer, support vector regression, backpropagation, long-short term memory**

I. Introduction

The cancer is called a malignant tumor, which is caused by an irregular division of any tissue or organ in cells [1]. When the treatment is not possible, it causes serious discomfort or death. From the past to the present, cancer cases have always been seen and continue to be seen in the future. For early diagnosis and the treatment of cancer diseases, studies and experiments have been carried out [2]-[5]. Although cancer cases can be diagnosed more quickly with the help of developing technology, it may cause death in cancer cases which can grow rapidly or metastasis to different regions.

There are many different types of cancer and this needs wide range of work to be studied. But, in spite of these studies, there are delays in diagnosis and treatment that causes losses because of cancer cases worldwide.

Although several types of cancer in both male and female exist, the most epidemic types of cancer in the world are lung, breast, prostate, stomach, liver, thyroid, ovary, esophagus, leukemia, pancreas [6].

There are a total of 184 countries in the world that have cancer registry and 28 cancer predictions that shows the increments of recent cancer incidences and cancer-related deaths [7]. Beside these, new cancer cases in the world are increased and became 14.1 million. In addition, the number of deaths related to cancer disease was 8.2 million. Some of the most diagnosed cancer types in the world are lung cancer (13.0%), breast cancer (11.9%) and colon cancer (9.7%) [7]. Most of the deaths are occurred by lung cancer (19.4%), liver cancer (9.1%) and gastric cancer (8.8%) as stated that cancer-related deaths occurred.

The report of World Health Organization [8] mentioned that the total number of new cases of lung cancer, which is considered in our study, is 18,078,957 million considering all age groups and both genders in the world. In addition, the number of deaths related to this cancer type is 9,555,027 million.

While the lung cancer has the highest mortality rate and the number of new cases [9], we will focus on lung cancer in our study. Therefore, we will examine the lung data in more details.

The comparison of previous and final reports shows that the total number of new lung cancer cases in the world has increased by about 4 million in 6 years and the number of deaths due to lung cancer, increased 1.3 million.

As it is common problem for our world, the most developed continent Europe has also struggle with this problem. Lung cancer rates of Europe are as follows [7]:

- Incidence for Lung Cancer -Both Sexes: 252,746 (12.1%).
- Mortality for Lung Cancer - Both Sexes: 173,278 (9.8%).

However, the starting years of the records of cancer incidences and mortality differs for each country and this makes predictions more difficult and the reliability of results becomes unstable. Thus, in this research, starting year is selected as 1970 which only 10 countries or regions have records at that year in Europe. These countries are Germany, Denmark, Estonia, Finland, Iceland, Norway, Slovakia, Slovenia, Sweden and Geneva region of Switzerland.

Artificial Intelligence and Machine Learning algorithms have gained an importance since last two decades to assist human-beings in order to improve the ability for analyzing the unstable data and to make stable decisions on them. They have been implemented almost in the every field of our lives. Machine learning algorithms have been developed to classify, predict or minimize the raw data. Since the proposed methods are not suitable for all kind of applications, several algorithms proved their ability in the prediction of data. Most common prediction algorithms are Support Vector Regression (SVR) [10], [11], Long Short Term Memory (LSTM) Network [12],

Manuscript received September 1, 2019; revised April 7, 2020.

doi: 10.12720/jait.11.2.91-96

Backpropagation Neural Network [13], [14], Radial Basis Function [15], Linear Regression [16] etc.

Kourou *et al.* [17] investigated several ML algorithms in order to determine the efficiency of machine learning techniques in cancer prognosis and prediction. They concluded that the researches are focused on supervised models for the development of predictive algorithms.

Implementation of ML algorithms in bioinformatics has gained an importance. Malvezzi *et al.* [18] used a Linear Regression model for the prediction of cancer mortality rates for the European Union and Ribes *et al.* [19] used Bayesian models to predict both incidence and mortality rates in Catalonia.

Alhaj and Maghari [20] implemented Random Forest and Rule Induction Algorithms to predict the cancer survivability rates in the Gaza Strip.

Recently, Jung *et al.* [21] used a Jointpoint regression model to predict cancer incidence and mortality rates in Korea for 2019.

In this paper, we predicted the cancer incidence rates of ten European countries mentioned above, starting from 1970 to 2012 using machine learning models, Support Vector Regression, Long-Short Term Memory Network and Backpropagation Neural Network.

The rest of the paper is organized as follows: Section 2 introduces the considered machine learning algorithms and Section 3 explains the design of experiments. Section 4 and Section 5 presents the results and discussions, respectively. Finally, Section 5 concludes the obtained results of this research and mentions the future works.

II. Machine Learning Algorithms

This section gives brief introduction about Machine Learning (ML) algorithms considered in this research namely Backpropagation Learning Algorithm (BPLA), Long-Short Term Memory (LSTM) and Support Vector Regression (SVR).

Several machine learning algorithms had been proposed. Some of them are based on statistical analysis of data and some of are neural-based models. Each algorithm had proved its efficiency in different kinds of applications and dataset. It is obvious that the efficiency and the success of the algorithms are strongly data-dependent. Therefore, the use of multiple ML algorithms and the comparative evaluation of considered algorithms are required for all kinds of applications. In this research, we have implemented three ML algorithms to perform comparative study and to determine the optimum for considered data.

A. Backpropagation Learning Algorithm

Backpropagation is a learning algorithm for multi-layer perceptron that updates weights of each neuron using gradient descent algorithm. Initial weights are generally randomly assigned and it starts by feeding inputs to the net and calculating total potential of following hidden layer by corresponding weights.

Activation function produces the output of each neuron and same calculations are repeated until output layer. At that layer, actual outputs are compared by targets and error is calculated. According to these error values, weights are updated until the convergence of neural network.

Backpropagation learning algorithm was used in several real-life applications in classification, prediction and optimization problems [14].

Fig. 1 shows general architecture of Backpropagation neural networks.

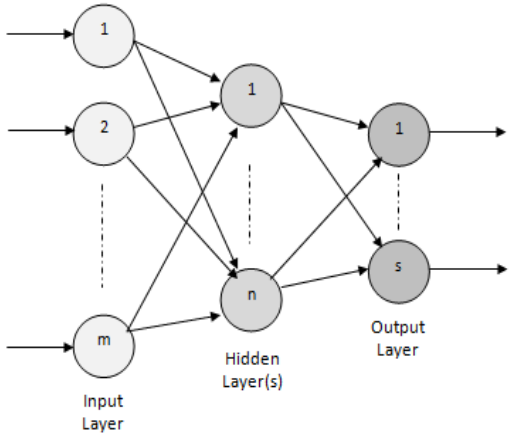


Figure 1. General architecture of backpropagation neural networks

B. Long-Short Term Memory Network

LSTM is an effective special version of recurrent network and generally used for classification and prediction problems [22], [23]. Four major components are formed its architecture: cell, input gate, output gate and forget gate. Forget gate is used for the removal of irrelevant data and input gate accepts the data form forget gate. Output gate produces the output of the LSTM cell using Sigmoid activation function. It uses gradients to update weights however, it remembers previous errors and this improves the error minimization of network in a short time.

Fig. 2 shows general architecture of LSTM.

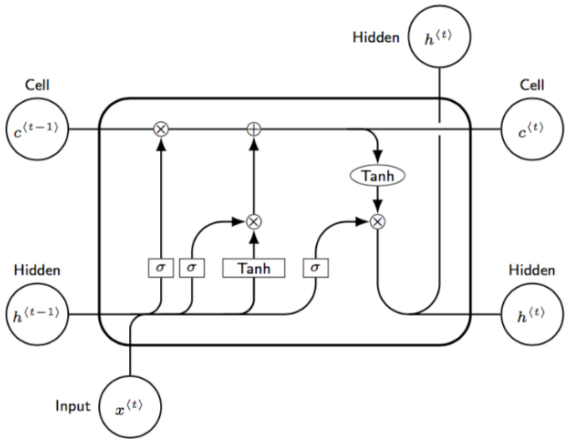


Figure 2. General architecture of LSTM (Image courtesy of stackexchange.com)

C. Support Vector Regression

Support Vector Regression is a kind of Support Vector Machines with a few changes to accept real value outputs instead of binary numbers. It is effectively used in prediction problems [24], [25]. It minimizes error by

maximizing the margin of hyper-plane. It creates the sub-class from training data which is called support vectors and tries to minimize the distance between the observed data and predicted data in order to improve the performance.

Fig. 3 shows general architecture of SVR.

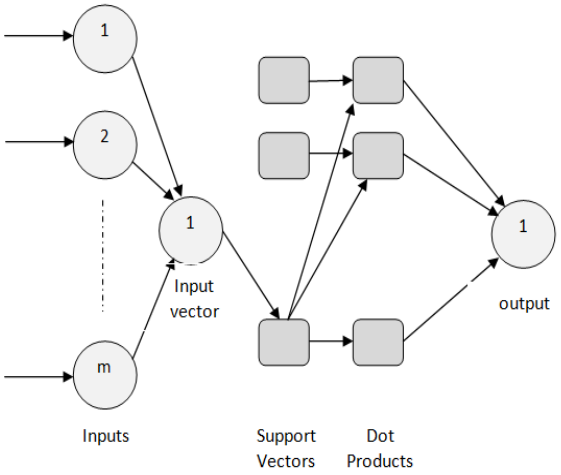


Figure 3. General architecture of SVR

III. Datasets and Design of Experiments

The 2012 reports of World Health Organization and Globacan [8] was used for Lung Cancer prediction for male and female separately. The data for 10 countries and 42 years were considered in experiments. However, the registration of cancer incidence records of the countries is not started to from the beginning of record dates, the initial start date of the records was decided to be used from 1970 in order to minimize missing data and improve prediction accuracy. The minimized number of missing values was replaced with the data imputation technique, the nearest neighbor value.

Data normalization was performed by a Min-Max normalizer to normalize data between 0 and 1 for each attribute. The formula of Min-Max normalizer is given in (1).

$$Z_i = \frac{(x_i - \min(x))}{\max(x) - \min(x)} \quad (1)$$

where x_i represents the real sample and, *min(x)* and *max(x)* denotes the minimum and maximum of the corresponding attribute.

Experiments were divided into two categories according to the gender. Then, each category was divided into sub-categories by considering training and testing samples sizes. Both of the sub-groups consists 60% and 70% of training ratio.

Totally 12 experiments were performed in order to analyze the reliability of the prediction results by considering different number of training and testing data and to achieve superior results by different algorithms.

Evaluation was performed according to 3 metrics, Mean Squared Error (MSE), R^2 Score and Explained Variance (EV) Score which are the main indicators of the success of predicted results and models.

Mean Squared Error calculates the squares of error of estimator and it is defined as:

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

where n is the total number of samples and Y_i and $\hat{Y}_i$ are the predicted and expected outputs of estimator respectively.

Explained Variance Score is another evaluation criteria of an estimator and also known as the regression sum of squares. It is defined as:

$$EV_s = \sum_{i=1}(f_i - \hat{y})^2 \quad (3)$$

where f_i is the predicted values and $\hat{y}$ is real sample.

R^2 Score is variance of predictable sample from the independent sample. It is defined as:

$$R^2 = \frac{EV_s}{UV_s} \quad (4)$$

where *EV* is defined in Equation 3 and *UV* is unexplained variations of samples.

Fig. 4 demonstrates the general block diagram of the performed experiments.

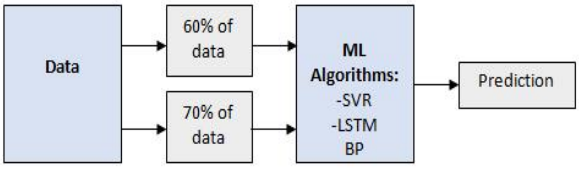


Figure 4. General block diagram of the performed experiments

IV. Experimental Results

In this section, results of performed experiments will be presented in details. As it was mentioned above, 12 experiments were performed in two groups for male and female, and two sub-groups for 60% and 70% of training data respectively. SVR, BP and LSTM network were used for the prediction.

Five-layered topology was used with Sigmoid activation function for both LSTM and Backpropagation neural networks. Radial Basis Function Kernel was used in SVR and γ and ε values were used as 0.005 and 0.01 respectively.

After several experiments optimum results were obtained after 350 epochs in Backpropagation, and 250 epochs in LSTM. Following subsections presents the obtained results for both male and female data.

A. Experimental Results for Male Group

For Male data, SVR produced superior results than Backpropagation and LSTM for all training samples for the prediction of lung cancer incidence with all indicators. *MSE*, R^2 and *EV* scores were obtained 0.00019, 0.9977 and 0.9980 respectively for 60% of training data and 0.0001341, 0.9999 and 0.9999 for 70% of training data which were the highest scores.

Table I shows the details of the results for Male Group with 60% and 70% of training samples respectively. Fig. 5 shows the prediction graph for SVR for both training sets of Male data.

TABLE I. RESULTS OF MALE GROUP PREDICTION

60% Training			
Result	SVR	LSTM	Backpropagation
MSE	**0.000196**	0.00701	0.00566
R^2	**0.99770**	0.91800	0.93300
EV	**0.9980**	0.93200	0.93700
70% Training			
Result	SVR	LSTM	Backpropagation
MSE	**0.0001341**	0.00600	0.00440
R^2	**0.9999**	0.92270	0.9423
EV	**0.9999**	0.92600	0.9440

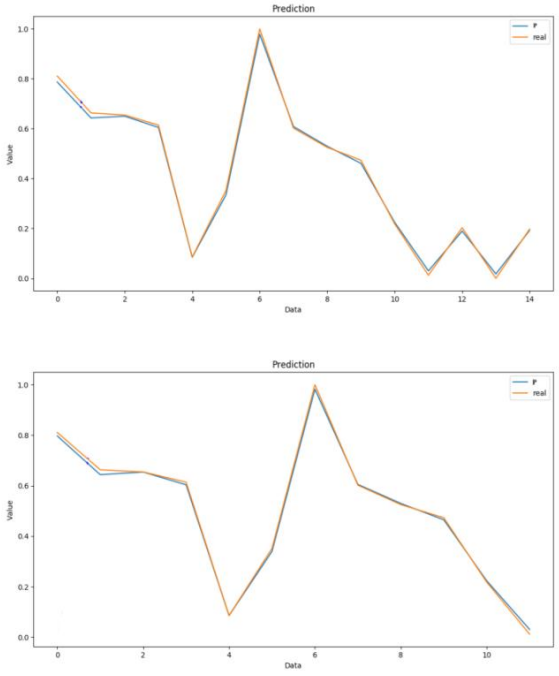


Figure 5. Prediction graphs of SVR for both training sets in Male Group

B. *Experimental Results for Female Group*

For Female data, similar to other experiments, SVR produced superior results for the prediction of lung cancer incidence with all indicators. *MSE*, R^2 and *EV* scores were obtained 0.00023236, 0.997002 and 0.99700 respectively for 60% of training data and 0.0000661, 0.998890 and 0.99990 for 70% of training data which were the highest scores.

Table II shows the details of the results for Female Group with 60% and 70% of training samples respectively. Fig. 6 shows the prediction graph for SVR for both training sets of Female data.

TABLE II. RESULTS OF FEMALE GROUP PREDICTION

	60% Training		
Result	SVR	LSTM	Backpropagation
MSE	**0.00023236**	0.006905	0.002170
R^2	**0.997002**	0.91090	0.97200
EV	**0.99700**	0.93800	0.97300
	70% Training		
Result	SVR	LSTM	Backpropagation
MSE	**0.00006612**	0.00234	0.00243
R^2	**0.998890**	0.96079	0.95930
EV	**0.99990**	0.97300	0.97300

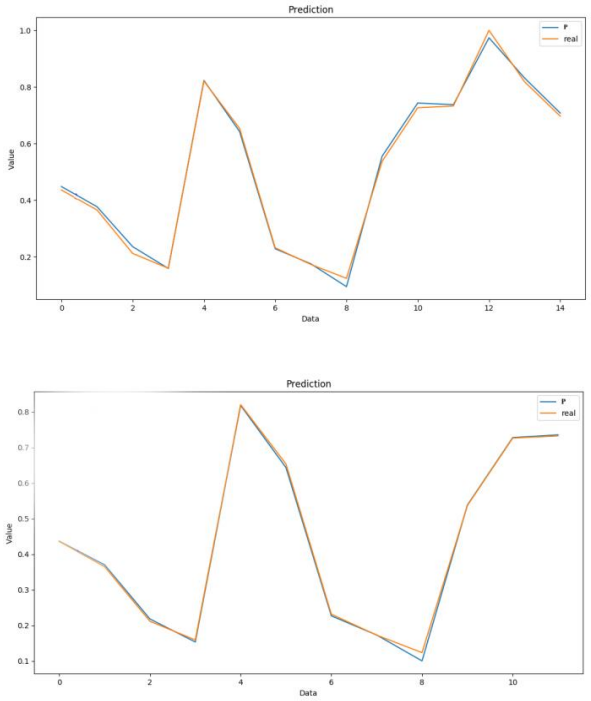


Figure 6. Prediction graphs of SVR for both training sets in Female Group

Even though superior results were obtained by SVR, when the comparison is performed between LSTM and Backpropagation, it can be seen that Backpropagation achieved relatively better results in prediction of lung cancer incidences than LSTM. Fig. 7 shows the graphs of best prediction results of Backpropagation and LSTM for both groups.

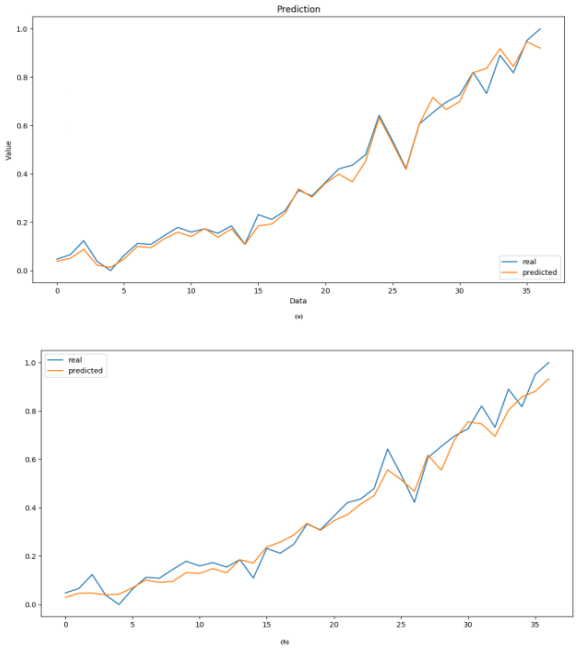
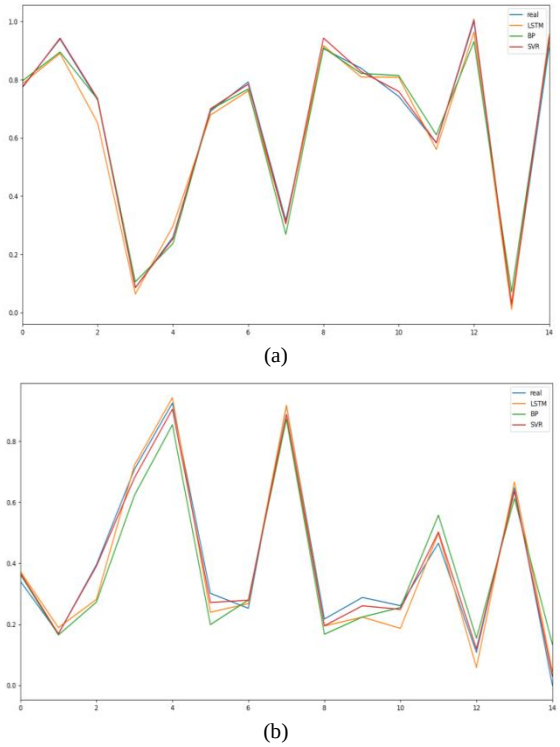


Figure 7. Best prediction graphs (a) backpropagation for 70% training of Female Group and (b) LSTM for 70% training of Female Group

V. Discussions

The results obtained in the experiments should be analyzed in two different ways, as the performance of considered models and the effect of training ratio on prediction performances of models.

In Male group experiments, it was observed that the use of 70% of total data for training increases the prediction performance of machine learning models. However, it was observed that the increment of training ratio in Female group causes little decrease in Backpropagation when R^2 and EV scores are considered. Also, increased MSE value was observed in Backpropagation with 70% of training ratio. Fig. 8 presents the prediction graphs of LSTM, Backpropagation and SVR for female and male groupsusing 60% of training ratio. But, similar to male group experiments, the performances of SVR and LSTM improved almost in all performance indicators when 70% of training ratio was considered.



(a)

(b)

Figure 8. Prediction graphs for SVR, LSTM and BP using 60% of training ratio, (a) Female data (b) Male data.

Therefore, it can be concluded that the increment of training data would minimize the error between predicted and observed data, thus increase the performance of the models to perform predictions.

The analyzes of obtained results show that machine learning models are suitable to predict untrained cancer incidence rates with high scores which would be used for the prediction of future rates and provide public awareness for cancer types.

VI. Conclusion

Cancer disease has a huge incidence and mortality rate worldwide. Reliable and steady data is not available because insufficient records. Lung cancer has highest mortality rates and this makes it more important to analyze the available data either it is insufficient.

Prediction of this kind of data is one of the most challenging tasks in machine learning and suitable algorithms should be selected to perform it.

In this paper, lung cancer incidence rates of male and female data for ten European countries were analyzed and prediction was performed using Support Vector Regression, Backpropagation and Long-Short Term Memory Network. Prediction results are analyzed by using most efficient evaluation criteria in the literature; *MSE*, R^2 and *EV* scores. Successful results were obtained for all algorithms; however, Support Vector Regression performed outstanding prediction results with the minimum error and the maximum prediction results. By considering other two algorithms, it was followed by Backpropagation and LSTM respectively.

Future work will include the implementation of more machine learning algorithms for the prediction of more cancer types for all European countries and dividing this predictions into age groups will considered to analyze the incidence rates for age groups. Also, mortality rates will be included to predict both incidence and the mortality rates of patients.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Kubra Tuncal conducted the research; Kubra Tuncal and BoranSekeroglu analyzed the data and performed simulations; Kubra Tuncal, Boran Sekeroglu and Cagri Ozkan wrote and revised the paper; all authors had approved the final version.

References

[1] M. A. M. Alhaj and A. Y. A. Maghari, "Cancer survivability prediction using random forest and rule induction algorithms," in *Proc. 8th International Conference on Information Technology*, 2017, pp. 388-391.

[2] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," in *Proc IEEMA Engineer Infinite Conference*, 2018, pp. 1-5.

[3] B. Bektas and S. Babur, "Machine learning based performance development for diagnosis of breast cancer," in *Proc. Medical Technologies National Congress (TIPTEKNO)*, 2016, pp. 1-4.

[4] D. Arslan, M. E. Auzdemir, and M. T. Arslan, "Diagnosis of pancreatic cancer by pattern recognition methods using gene expression profiles," in *Proc. International Artificial Intelligence and Data Processing Symposium*, 2017, pp. 1-4.

[5] E. Razak, F. Yusof, and R. A. Raus, "Classification of mirna expression data using random forests for cancer diagnosis," in *Proc. International Conference on Computer and Communication Engineering*, 2016, pp. 187-190.

[6] World Health Organization. United Nations. (2018). [Online]. Available: www.who.org

[7] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistic, 2012," *A Cancer Journal for Clinicians*, vol. 65, pp. 87-108, 2015.

[8] F. Bray, M. Colombet, L. Mery, M. Pieros, A. Znaor, R. Zanetti, and J. Ferlay. (2018). Cancer incidence in five continents. [Online]. Available: http://ci5.iarc.fr last accessed on november

[9] A. Srinidhi S. Lam A. McWilliams, P. Beigi, and C. E. MacAulay, "Sex and smoking status effects on the early detection of early lung

cancer in high-risk smokers using an electronic nose," *IEEE Trans. on Biomedical Engineering*, vol. 62, no. 8, pp. 2044-2054, August 2015.
[10] J. Zou, C. Li, Q. Yang, and Q. Li, "Fault prediction method based on SVR of improved PSO," in *Proc. the 27th Chinese Control and Decision Conference*, 2015, pp. 1671-1675.
[11] Z. Zhang, X. Wang, and Y. Ji, "The power load forecasting of SVR based on Hadoop," in *Proc. 37th Chinese Control Conference*, 2018, pp. 4484-4488.
[12] Z. Chen, Y. Liu, and S. Liu, "Mechanical state prediction based on LSTM neural network," in *Proc. 36th Chinese Control Conference*, 2017, pp. 3876-3881.
[13] A. Khashman and B. Sekeroglu, "Document image binarisation using a supervised neural network," *International Journal of Neural Systems*, vol. 18, pp. 405-418, 2008.
[14] T. Adali and B. Sekeroglu, "Analysis of MicroRNAs by neural network for early detection of cancer," *Procedia Technology Elsevier*, vol. 1, pp. 449-452, 2012.
[15] F. J. Chang, J. M. Liang, and Y. C. Chen, "Flood forecasting using radial basis function neural networks," *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 31, pp. 530-535, 2001.
[16] F. S. Gharehchopogh, T. H. Bonab, and S. R. Khaze, "A linear regression approach to prediction tf Stock market trading volume: A case study," *International Journal of Managing Value and Supply Chains*, vol. 4, no. 3, 2013.
[17] K. Kourou, T. P. Exarchos, and K. P. Exarchos, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8-17, 2015.
[18] M. Malvezzi, P. Bertuccio, and F. Levi, "European cancer mortality predictions for the year 2014," *Annals of Oncology*, vol. 25, pp. 1650-1656, 2014.
[19] J. Ribes, L. Esteban, and R. Clries, "Cancer incidence and mortality projections up to 2020 in Catalonia by means of Bayesian models," *Clinical and Translational Oncology*, vol. 16, pp. 714-724, 2014.
[20] M. A. M. Alhaj and A. Y. A. Maghari, "Cancer survivability prediction using random forest and rule induction algorithms," in *Proc. 8th International Conference on Information Technology*, 2017.
[21] K. W. Jung, Y. J. Won, H. J. Kong, and E. Lee, "Prediction of cancer incidence and mortality in Korea," *Cancer Research and Treatment*, vol. 51, no. 2, pp. 431-437, 2019.
[22] S. Li, Q. Wang, X. Liu, and J. Chen, "Low cost LSTM Implementation based on stochastic computing for channel state information prediction," in *Proc. IEEE Asia Pacific Conference on Circuits and Systems*, 2018, pp. 231-234.
[23] S. Dai, L. Li, and Z. Li, "Modeling vehicle interactions via modified LSTM models for trajectory prediction," *IEEE Access*, vol. 7, pp. 38287-38296, 2019.
[24] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student performance prediction and classification using machine learning models," in *Proc. 8th International Conference on Educational and Information Technology*, 2019.
[25] L. Ge, J. Shi, and P. Zhu, "Melt index prediction by support vector regression," in *Proc. International Conference on Control, Automation and Information Sciences*, Ansan, 2016, pp. 60-63.

Kubra Tuncal was born in Malatya, Turkey on 20th of July 1987. She completed Inonu University Akcadag Vocational High School at the Department of Computer Technology and Programming, in 2006. After completing associate degree, in 2009, she started to undergraduate study at Near East University in Information Systems Engineering. She obtained BSc. degree in 2017 with honor degree. She obtained MSc. degree at the Department of Information Systems Engineering in 2019 and she has been working as a Research Assistant in same department since 2017.

Boran Sekeroglu was born in Nicosia, Cyprus on 21st of December 1980. He completed BSc., MSc. and PhD. programs in Department of Computer Engineering of Near East University with honors. After obtaining PhD. degree, he completed his one-year military service. He worked as a Vice Chairman in Computer Engineering Department between 2009 and 2013, and he has been working as a teaching staff and chairman at Department of Information Systems Engineering since 2013as Assistant Professor.
He published more than 40 SCI articles, conference papers, book chapters and books. He made several contributions to Journals, Conferences as a science board member, reviewer and consultant. He supervised more than 20 undergraduate graduation projects and 20 MSc. thesis.

Cagri Ozkan was born in Alanya, Turkey on 17th of September 1990. He completed Mehmet Akif Ersoy University H.T. Vocational High School at the Department of Computer Technology and Programming in 2010. After completing associate degree, he started to undergraduate study at Near East University in Information Systems Engineering. He obtained BSc. degree in 2015 as a first ranking student and he started MSc. education in same year. He obtained MSc. degree at the Department of Information Systems Engineering in 2017. He has been working as a teaching staff in same department of Near East University.