



THE JOURNAL OF

COGNITIVE SYSTEMS

an international, peer-reviewed  
indexed, and open-access periodical

VOLUME 07  
NUMBER 02  
YEAR 2022  
ISSN 2548-0650

[www.dergipark.gov.tr/jcs](http://www.dergipark.gov.tr/jcs)





# THE JOURNAL OF COGNITIVE SYSTEMS

An international, peer-reviewed,  
indexed, and open-access periodical

ISSN 2548-0650

## GENERAL PUBLICATION DIRECTOR & OWNER

ISMAIL KOYUNCU, (RECTOR), Istanbul Technical University, Turkey

## EDITOR-IN-CHIEF

S. SERHAT SEKER, Istanbul Technical University, Turkey

## EDITORIAL BOARD

CEMIL COLAK, Inonu University, Turkey  
EVREN DAGLARLI, Istanbul Technical University, Turkey  
GOKHAN ERDEMIR, University of Tennessee at Chattanooga, USA  
T. CETIN AKINCI, Istanbul Technical University, Turkey  
ONUR ERGEN, Istanbul Technical University, Turkey  
ZEYNEP KUCUKAKCALI, Inonu University, Turkey

## SCIENTIFIC COMMITTEE

ABRAHAM LOMI (Indonesia)  
ALTAN CAKIR (Turkey)  
AYKUT HOCANIN (Turkish Republic of Northern Cyprus)  
BELLE R. UPADHYAYA (USA)  
BERK USTUNDAG (Turkey)  
BURAK BARUTCU (Turkey)  
BURCU OZDEMIR (Turkey)  
CHANDAN KUMAR CHANDA (India)  
DENIZ TURKPENCE (Turkey)  
ERKAN KAPLANOGLU (USA)  
ESEN YILDIRIM (Turkey)  
GOKHAN ERDEMIR (Turkey)  
HAKAN TEMELTAS (Turkey)  
HASAN DEMIREL (Turkish Republic of Northern Cyprus)  
JELENA DIKUN (Lithuania)  
KUNIHICO NABESHIMA (Japan)  
MURAT OKATAN (Turkey)  
MUSA YILMAZ (Turkey)  
NECDET OSAM (Turkish Republic of Northern Cyprus)  
OKYAY KAYNAK (Turkey)  
OLEKSII TURUTA (Ukraine)  
OSMAN NURI UCAN (Turkey)  
OMER FARUK ERTUGRUL (Turkey)  
RITUPARNA DATTA (South Korea)  
SALIH BARIS OZTURK (Turkey)  
TANJU SURMELI (Turkey)  
UFUK KORKMAZ (Turkey)

## AIM & SCOPE

The Journal publishes original papers in the field of artificial intelligence, biomedical, quantum information, big data analysis and statistical areas, which are related with cognitive science and engineering, as well as the applications of the cognitive studies in social areas. Letter to the editor is also encouraged.

## THE JOURNAL OF COGNITIVE SYSTEMS

THE JOURNAL OF COGNITIVE SYSTEMS (JCS) is published bi-annually. Responsibility for the content rests upon the authors and not upon the JCS or the publisher. **Reuse Rights and Reprint Permissions:** Educational or personal use of this material is permitted without fee, provided such use: i) is not made for profit; and ii) includes this notice and a full citation to the original work on the first page of the copy; and iii) does not imply JCS endorsement of any third-party products or services. Authors and their companies are permitted to post their JCS-copyrighted material on their own web servers without permission, provided that the JCS copyright notice and a full citation to the original work appear on the first screen of the posted copy. Permission to reprint/republish this material for commercial, advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from JCS by writing to the JCS General Publication Director Prof. Dr. S. Serhat Seker, Faculty of Electrical and Electronics Engineering, Istanbul Technical University, Istanbul, Turkey; or [sekers@itu.edu.tr](mailto:sekers@itu.edu.tr).  
Copyright © 2016 JCS. All rights reserved.

## CONTACT

Prof. Dr. Ş. Serhat Seker  
Editor-in-Chief of The Journal of Cognitive Systems  
Delft University of Technology, The Netherlands  
Istanbul Technical University, Istanbul, Turkey

## E-MAIL

[cognitive@itu.edu.tr](mailto:cognitive@itu.edu.tr)

## WEB PAGE

[www.cognitive.itu.edu.tr](http://www.cognitive.itu.edu.tr)

## URL

[www.dergipark.org.tr/jcs](http://www.dergipark.org.tr/jcs)





# THE JOURNAL OF COGNITIVE SYSTEMS

VOLUME 07, NUMBER 02

DECEMBER 2022

## CONTENTS

<b>O. Ozhan, and F.H.Yagin:</b> Machine Learning Approach for Classification of Prostate Cancer Based on Clinical Biomarkers, .....	17-20
<b>O. Ozhan, and Z. Kucukakcali :</b> Risk Prediction Model for Dementia by Deep Learning Using Clinical Data, .....	21-24
<b>S. Yasar, and S. Yologlu :</b> Prediction of Placenta Accreta Spectrum by Machine Learning Methods and Determination of Candidate Biomarkers, .....	25-28
<b>F. Inceoglu, and F.H. Yagin :</b> Genomic Biomarkers of Metastasis in Breast Cancer Patients: A Machine Learning Approach,.....	29-32
<b>O. Topsakal, and E. Topsakal:</b> Framework for A Foreign Language Teaching Software for Children Utilizing AR, Voicebots and ChatGPT (Large Language Models) , .....	33-38



# Machine Learning Approach for Classification of Prostate Cancer Based on Clinical Biomarkers

Onural Ozhan<sup>1</sup>, Fatma Hilal Yagin<sup>2</sup>

<sup>1</sup> Department of Medicinal Pharmacology, Faculty of Medicine, İnönü University, Malatya, 44210, Turkey (e-mail: [onural.ozhan@inonu.edu.tr](mailto:onural.ozhan@inonu.edu.tr)).

<sup>2</sup> Department of Biostatistics, and Medical Informatics, Faculty of Medicine, İnönü University, Malatya 44210, Turkey (e-mail: [hilal.yagin@inonu.edu.tr](mailto:hilal.yagin@inonu.edu.tr)).

## ARTICLE INFO

**Received:** Aug.,03.2022

**Revised:** Oct, 11.2022

**Accepted:** Dec.,21.2022

### Keywords:

Machine learning  
Classification  
Prostate cancer  
Random forest

**Corresponding author:** Onural Özhan

✉ [onural.ozhan@inonu.edu.tr](mailto:onural.ozhan@inonu.edu.tr)

☎ +90 535 932 33 44

**ISSN:** 2548-0650

**DOI:**

<https://doi.org/10.52876/jcs.1221425>

## ABSTRACT

In this study, it is aimed to classify cancer based on machine learning (ML) and to determine the most important risk factors by using risk factors for prostate cancer patients. Clinical data of 100 patients with prostate cancer were used. A prediction model was created with the random forest (RF) algorithm to classify prostate cancer. The performance of the model was obtained by Monte-Carlo cross validation (MCCV) using balanced subsampling. In each MCCV, two-thirds (2/3) of the samples were used to assess the significance of the feature. In order to evaluate the performance of the model, graph, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-score and Area under the ROC Curve (AUC) criteria including prediction class probabilities and confusion matrix were calculated. When the results were examined, the sensitivity, specificity, positive predictive value, negative predictive value, accuracy, F1-score, and AUC values obtained from the RF model were 0.89, 0.84, 0.77, 0.93, 0.86, 0.83, and 0.88, respectively. Area, perimeter, and texture were the three most important risk factors for differentiating prostate cancer. In conclusion, when the RF algorithm can be successfully predicted prostate cancer. The important risk factors determined by the RF model may contribute to diagnosis, follow-up and treatment researches in prostate cancer patients.

## 1. INTRODUCTION

THE abnormal division of cells in the prostate gland is one of the characteristics that define the type of cancer known as prostate cancer [1]. Research indicates that prostate cancer is the second most common form of cancer in men and the fifth leading cause of death on a global scale. On the other hand, it is the type of cancer that is diagnosed in more men over the age of middle age than any other type, in both developed and developing nations. If a man is between the ages of 40 and 59, his risk of developing prostate cancer is 2.58%, but between the ages of 60 and 79, his risk increases to 14.7%. The probability of developing prostate cancer in a man between the ages of 0 and 39 is only 0.01%. In addition, the likelihood of a man developing prostate cancer during the course of his lifetime is approximately 17.8% [2,3].

The predisposition of one's family to develop prostate cancer, as well as advanced age, race, genetics, diet, environmental factors, and hormonal factors, are all considered to be risk factors [4]. The correct management of the treatment, diagnosis, and follow-up process of prostate

cancer is important not only for the patient and the doctor, but also for national health policies [5,6].

For this reason, the need for methods that can detect prostate cancer rapidly and accurately is increasing.

Machine learning, also known as ML, is a subset of artificial intelligence that identifies patterns in unprocessed data through the application of a specific algorithm or method. The primary objective of machine learning is to make it possible for computer systems to learn from experience on their own, without the need for explicit programming or intervention from humans. ML methods are frequently used in different areas of medicine and are less costly, more accurate and faster results in the diagnosis of different diseases. ML methods increase the predictive power thanks to their ability to combine data from various sources and manage large amounts of data [7-11].

Classification is one of the important tasks of ML. Classification includes approaches used to estimate the output variable when the output variable is categorical. The model obtained by using classification algorithms is used to predict

the unknown output variable when new data is obtained [12-14].

In this study, it is aimed to classify cancer based on ML and to determine the most important risk factors by using risk factors for prostate cancer patients.

## 2. MATERIAL AND METHODS

### 2.1. Data

The dataset to classify and predict prostate cancer in the study was obtained from <https://www.kaggle.com/sajidsaifi/prostate-cancer>. Of the patients in the data set, 62 (62%) were diagnosed as malignant and 38 (38%) as benign. Variables used to predict prostate cancer in the dataset: radius (mean distances from the center to perimeter points), texture (the standard deviation of grayscale values), perimeter (mean size of the core tumor), area, smoothness (mean of local variation in radius lengths), compactness ((mean of perimeter)<sup>2</sup> / (area - 1)), symmetry and fractal dimension (mean for "coastline approximation").

### 2.2. Methods

#### 2.2.1. Machine Learning Approach

The Random forest (RF) algorithm was used to predict prostate cancer in the study. The RF algorithm is among the ensemble classification methods created by Leo Breiman. Ensemble classification techniques are learning algorithms that produce more than one classifier instead of just one classifier and classify new data with the votes obtained as a result of the predictions of the classifiers produced [15,16].

In the RF classification method, as in other ensemble learning methods, the performance values of weak learners (single decision tree, single sensor, etc.) are increased by a voting scheme. The classification method with the RF algorithm is based on the decision tree model. One of the advantages of the random forest algorithm is that it can use both continuous and discrete variables together. It can also be used in large or small size data sets [17,18].

The performance of the model was obtained by Monte-Carlo cross validation (MCCV) using balanced subsampling [19]. In each MCCV, two-thirds (2/3) of the samples were used to assess the significance of the feature. In order to evaluate the performance of the model, graph, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-score and Area under the ROC Curve (AUC) criteria including prediction class probabilities and confusion matrix were calculated.

## 3. RESULTS

In order to examine the performance of the RF model, confusion matrix with class probabilities is given in Figure 1. According to Figure 1, the model correctly classified (predicted) 52 of the 62 malignant patients and misclassified 10 patients.

Table 1 shows the results of the criteria related to the performance of the model, and Figure 2 shows the ROC curve. When the results were examined, the sensitivity, specificity, positive predictive value, negative predictive value, accuracy, F1 score, and AUC values obtained from the RF model were 0.89, 0.84, 0.77, 0.93, 0.86, 0.83, and 0.88, respectively

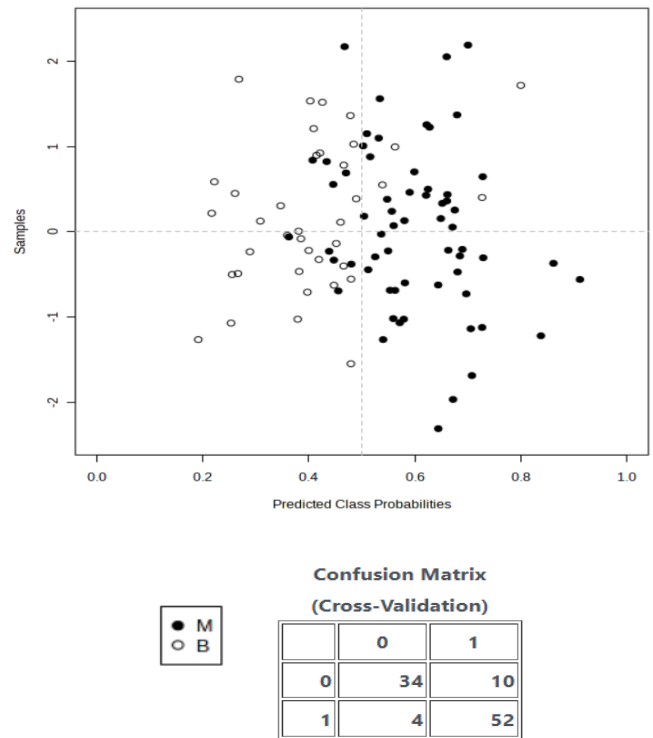


Fig. 1. Class probabilities and confusion matrix for model estimation (0: benign; 1 = malignant)

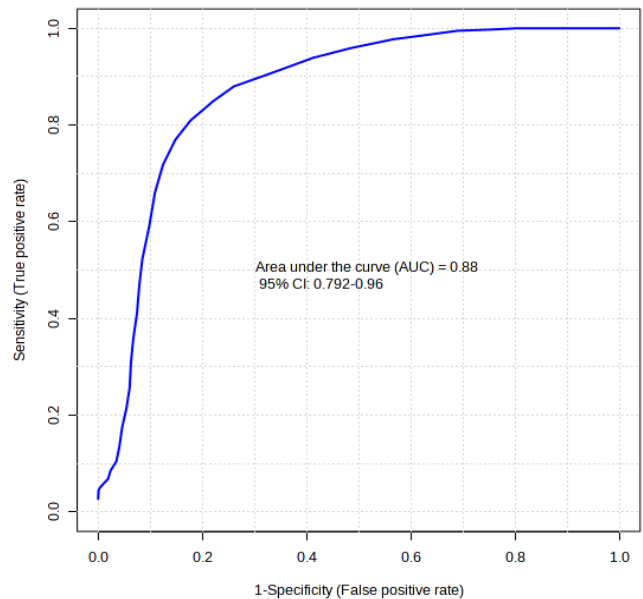


Fig. 2. Modelin performansına ilişkin ROC eğrisi

In Figure 3, the importance plot of the variables according to their contribution to the RF model created to predict prostate cancer is given. According to Figure 3, area, perimeter, and texture were the three most important factors to differentiate prostate cancer. In particular, the importance level of area was high compared to other risk factors.

TABLE I  
The Performance of the Models

Metric	Value
Sensitivity	0.89
Specificity	0.84
Positive Predictive Value	0.77
Negative Predictive Value	0.93
Accuracy	0.86
F1 score	0.83
AUC	0.88

AUC: Area under the ROC Curve

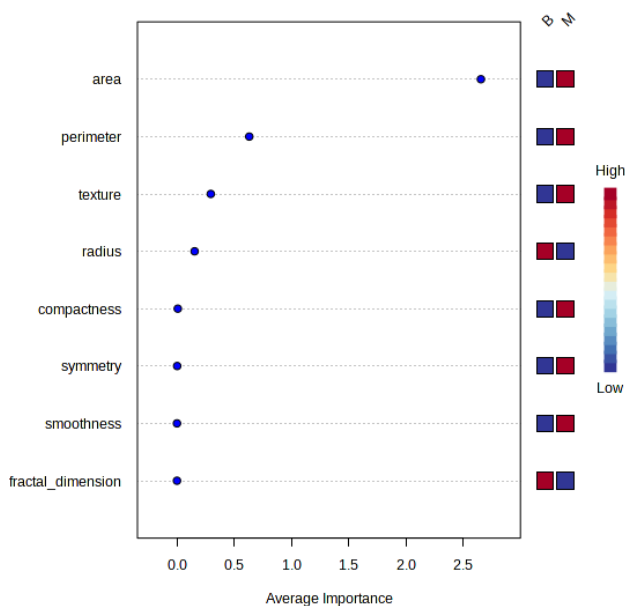


Fig. 3. Importance plot for clinical factors

#### 4. DISCUSSION

The prostate cancer is a disease that can begin in any part of the prostate gland, progress slowly for the first five to ten years, and then rapidly grow after that and can also spread to other organs. Cancer of the prostate is a significant contributor to male health problems and deaths. Initiating factors for prostate cancer, such as genetic factors, chronic inflammation and infection, high-fat diet, smoking, alcohol use, and obesity, are not fully understood at this time. Prostate cancer is caused by a combination of factors. In terms of both incidence and mortality, prostate cancer ranks among the top five most common cancers in the world. Therefore, early detection of prostate cancer allows for the possibility of preventing the progression of the disease as well as the application of alternative treatment protocols [20,21].

ML methods have been used frequently for cancer detection and classification in recent years. Clinical decision support systems developed based on ML can help clinicians in the pre-diagnosis, follow-up and treatment of diseases [22,23].

This research focused on the prediction of prostate cancer with ML methods, which is one of the most common causes of cancer-related death in men and shows symptoms similar to benign enlargement. Diagnosing diseases is one of the most challenging aspects of the medical field. The fact that there are

no established guidelines for evaluating prostate cancer symptoms and that the diagnostic methods that are currently available have poor predictive rates makes this study extremely valuable. In situations where there are no hard-and-fast rules to follow but where the factors that will influence an event can be anticipated, such as in the case of prostate cancer, we believe that methods of machine learning may be useful in making accurate predictions.

Based on this, the RF algorithm, which is one of the supervised machine learning methods and contributed to the creation of high-performance models, was used in this research to evaluate the accuracy of the prediction of prostate cancer.

In the model, patients were assigned to one of the output classes based on the class probabilities, thus obtaining the confusion matrix. The sensitivity, specificity, positive predictive value, negative predictive value, accuracy, F1 score, and AUC values obtained from the RF model were 0.89, 0.84, 0.77, 0.93, 0.86, 0.83, and 0.88, respectively. Our results showed that the RF model could successfully predict prostate cancer. In addition, the importance of the clinical features examined in order to distinguish prostate cancer was examined in the study. Our results showed that area, perimeter, and texture are the most important features in differentiating prostate cancer.

In a study using the same data set in the literature and comparing the results using various ML approaches, the highest classification accuracy was obtained with the k-nearest neighbor and naive bayes methods [24]. The classification rate accuracy obtained with the optimal model of this study was found to be 0.80. Another study using the same dataset compared the performance of some popular ML methods to predict prostate cancer. The authors achieved the best performance with the Recurrent Neural Network (RNN) model with an accuracy rate of 0.813 [25]. In our study, we classified prostate cancer with an accuracy of 0.86, and the RF model had the ability to discriminate quite well.

As a result, the proposed RF model can successfully classify prostate cancer and the model can help clinicians to pre-diagnose prostate cancer.

#### 5. CONCLUSIONS

In conclusion, prostate cancer risk can be successfully predicted with methodology combined using clinical information and RF algorithm. Furthermore, the model created with the RF algorithm can help clinicians in the diagnosis, follow-up and treatment of patients with prostate cancer.

#### REFERENCES




- [1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6), 394-424.
- [2] Jemal, A. (2005). murray t, Ward e, samuels A, tiwari RC, Ghafoor A, Feuer eJ, thun mJ. *Cancer statistics*, 10-30.
- [3] Rawla, P. (2019). Epidemiology of prostate cancer. *World journal of oncology*, 10(2), 63.
- [4] Jemal, A., Thomas, A., Murray, T., & Thun, M. (2002). *Cancer statistics, 2002. Ca-A Cancer Journal for Clinicians*, 52(1), 23-47.
- [5] Siegel, R. L., Miller, K. D., & Jemal, A. (2019). *Cancer statistics, 2019. CA: a cancer journal for clinicians*, 69(1), 7-34.
- [6] Dimakakos, A., Armakolas, A., & Koutsilieris, M. (2014). Novel tools for prostate cancer prognosis, diagnosis, and follow-up. *BioMed research international*, 2014.
- [7] Yağın, F. H., Yağın, B., Arslan, A. K., & Çolak, C. (2021). Comparison of Performances of Associative Classification Methods for Cervical

- Cancer Prediction: Observational Study. *Turkiye Klinikleri Journal of Biostatistics*, 13(3).
- [8] Deo RC. (2015). Machine learning in medicine: *Circulation*, 132(20), 1920-30.
- [9] Sidey-Gibbons, J. A., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction: *BMC medical research methodology*, 19(1), 1-18.
- [10] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction: *Computational and structural biotechnology journal*, 13, 8-17.
- [11] Richter, A. N., & Khoshgoftaar, T. M. (2018). A review of statistical and machine learning methods for modeling cancer risk using structured clinical data: *Artificial intelligence in medicine*, 90, 1-14.
- [12] Paksoy, N., & Yağın, F. H. (2022). Artificial Intelligence-based Colon Cancer Prediction by Identifying Genomic Biomarkers: *Medical Records*, 4(2), 196-202.
- [13] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques: *Artificial Intelligence Review*, 26(3), 159-190.
- [14] Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: applications and issues: *Journal of Basic & Applied Sciences*, 13, 459-465.
- [15] Yilmaz, R., & Yağın, F. H. (2022). Early detection of coronary heart disease based on machine learning methods: *Medical Records*, 4(1), 1-6.
- [16] Khan, M. A., Memon, S. A., Farooq, F., Javed, M. F., Aslam, F., & Alyousef, R. (2021). Compressive strength of fly-ash-based geopolymer concrete by gene expression programming and random forest: *Advances in Civil Engineering*, 2021.
- [17] Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model: *Big Data Mining and Analytics*, 4(2), 116-123.
- [18] Palimkar, P., Shaw, R. N., & Ghosh, A. (2022). Machine learning technique to prognosis diabetes disease: random forest classifier approach *Advanced Computing and Intelligent Technologies*: Springer, 219-244.
- [19] Shan, G. (2022). Monte Carlo cross-validation for a study with binary outcome and limited sample size: *BMC Medical Informatics and Decision Making*, 22(1), 1-15.
- [20] Gandaglia, G., Leni, R., Bray, F., Fleshner, N., Freedland, S. J., Kibel, A., . . . La Vecchia, C. (2021). Epidemiology and prevention of prostate cancer: *European urology oncology*.
- [21] Habib, A., Jaffar, G., Khalid, M. S., Hussain, Z., Zainab, S. W., Ashraf, Z., . . . Habib, P. (2021). Risk Factors Associated with Prostate Cancer: *Journal of Drug Delivery and Therapeutics*, 11(2), 188-193.
- [22] Yağın, F. H., Göldoğan, E., Ucuzal, H., & Çolak, C. (2021). A Computer-Assisted Diagnosis Tool for Classifying COVID-19 based on Chest X-Ray Images: *Konuralp Medical Journal*, 13(S1), 438-445.
- [23] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine: *New England Journal of Medicine*, 380(14), 1347-1358.
- [24] <https://www.kaggle.com/alihantabak/prostate-cancer-predictions-with-ml-and-dl-methods>.
- [25] Laabidi, A., & Aissaoui, M. (2020). Performance analysis of Machine learning classifiers for predicting diabetes and prostate cancer: Paper presented at the 2020 1st international conference on innovative research in applied science, engineering and technology (IRASET).
- Onural Ozhan** obtained his BSc. degree in pharmacy from Inonu University in 2009. He received MSc. degree in pharmacognosy from Ankara University in 2012. He received Ph.D. degree in medical pharmacology from Inonu University in 2019. In 2015, he joined the department of medical pharmacology at Inonu University as a research assistant. Currently, he continues to work as an Asst. Prof. at Inonu University. His research interests are the cardiovascular system, oxidative stress, and ischemia-reperfusion injury.
- Fatma Hilal Yagin** obtained her BSc. degree in Statistics from Gazi University in 2017. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2020. She currently continues Ph.D. education in biostatistics and medical informatics from the Inonu University. In 2019, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning, and image processing.

## BIOGRAPHIES



# Risk Prediction Model for Dementia by Deep Learning Using Clinical Data

<sup>1</sup>Onural ÖZHAN , <sup>2</sup>Zeynep Küçükakçalı , <sup>3</sup>İpek Balıkçı Çiçek 

<sup>1</sup> Inonu University, Faculty of Medicine, Department of Medical Pharmacology, Malatya, Turkey. (e-mail: onural.ozhan@inonu.edu.tr)

<sup>2,3</sup> Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey. (e-mail: { zeynep.tunc, ipek.balikci }@inonu.edu.tr).

## ARTICLE INFO

**Received:** July.,09.2022

**Revised:** Aug.,28.2022

**Accepted:** Sep.,17.2022

### Keywords:

Dementia,  
Classification,  
Machine learning,  
Deep learning,  
Variable importance.

**Corresponding author:** Onural Özhan

✉ onural.ozhan@inonu.edu.tr

☎ +90 535 932 33 44

**ISSN:** 2548-0650

**DOI:** <https://doi.org/10.52876/jcs.118828>

## ABSTRACT

It is estimated that dementia, which is the most important public health problem in the elderly, will increase day by day. It is stated that this situation will create great challenges for public health and aged care systems in all countries of the world. For this reason, it has become very important to determine the management and treatment procedures of dementia, to reduce the level of progression of the disease and to increase the quality of life of individuals exposed to the disease. The purpose of this study is to predict dementia and reveal the factors related to the disease with the deep learning approach.

In the current study, open-access dementia data, which includes the information of 376 patients, was used. Dementia prediction was made using the deep learning method. Model results were evaluated with accuracy, balanced accuracy, sensitivity, selectivity, positive predictive value, negative predictive value, and F1-score performance metrics. In addition, 10-fold cross-validation method was used in the modeling phase. Finally, variable importance values were obtained by modeling.

When the results are examined The highest metric values among the performance criteria achieved for group variable types were calculated for Demented; and were found that Accuracy, Sensitivity, Specificity, Positive predictive value, Negative predictive Value, and F1-score were 0.947, 0.946, 0.978, 0.966, 0.965 and 0.956 respectively.

As a result, when the findings obtained from this study were examined, the dementia dataset, which consisted of imaging data and information about patients with clinical data, was classified with high accuracy using the deep learning method. The risk factors for dementia were determined with the variable importance values obtained as a result of the model.

## 1. INTRODUCTION

**D**EMENTIA is one of the most important public health problems frequently seen in the elderly, and it occurs with chronic and usually progressive decrease in memory and intellectual capacity. It is characterized by impairment in at least two cognitive functions such as memory, speech, perception, calculation, orientation, judgment, abstract thinking, and problem solving [1,2].

Dementia is very common nowadays. In 2015, it was determined that there were 47.47 million people diagnosed with dementia in the world. It is predicted that the number of people diagnosed with dementia will increase day by day, and it is expected that there will be 75.63 million people diagnosed with dementia in 2030 and 135.46 million in 2050 [3].

In parallel with the rapid increase in the elderly population, diseases that are common in old age are becoming a problem in the society. Dementia disease, which is encountered as a major health problem in aging societies

and develops due to vascular-degenerative changes, follows an irreversible and progressive process. The main goals in the care of patients with dementia are to protect the patient's functionality, reduce his disability, regulate the environment and relationships to maintain stability, compensate for the losses caused by the disease, and provide a therapeutic environment that will protect the patient's individuality and sustain the quality of life. In order to achieve these goals, the risk factors of the disease should be known and the living conditions should be supported by stabilizing the patients. Therefore, studies are needed to determine the risk factors for the disease [4].

Deep learning is a form of machine learning that enables computers to learn from experience and understand the world in terms of a hierarchy of concepts. Deep learning is based on the principle that each sequential layer receives the output of the previous layer as its input [5]. It's a representation learning algorithm made up of a number of complex non-



linear transformation structures, as well as multiple intermediate processing layers, that can perform high-level abstraction on data automatically [6].

Deep neural networks, like shallow neural networks, are feed-forward networks with one or more hidden layers that can provide modeling for extremely complex non-linear problems [7]. Deep learning models are preferred to obtain results such as diagnosis, classification, estimation, and detection. The use of machine learning methods in the early diagnosis of diseases in the field of health has increased. Biomedical, health bioinformatics and medical imaging are among the areas where machine learning methods are most popular. It is seen that deep learning methods that are successful in analyzing big data with different network architectures and learning algorithms will help healthcare professionals both in the early diagnosis of diseases and in the early treatment of the disease [8].

The aim of this study is to use a deep learning approach on an open access dementia data set to identify dementia situations and evaluate relevant factors.

## 2. MATERIAL and METHODS

### 2.1. Data set

The open access "Dementia Classification" data set to be used in the study can be accessed at <https://www.kaggle.com/deepak525/dementia-classification-compare-classifiers>. There are 373 patients in the data set used. 146 of these patients are demented, 190 non-demented patients and 37 converted patients. The explanations of the variables are given in Table 1.

TABLE I  
THE EXPLANATIONS OF THE VARIABLES

Variables	Explanations of Variable
CDR	Clinical Dementia Rating
MMSE	Mini Mental State Examination
VISIT	Number of visit
ETIV	Estimated Total Intracranial Volume
SES	Socioeconomic Status
MRDELAY	““
MF	Sex
EDUC	Years of Education
ASF	Atlas Scaling Factor
NWBV	Normalize Whole Brain Volume
Age	-

### 2.2. Deep Learning

Deep learning is a subfield of machine learning that includes methods for learning representations at various levels on data with complex relationships. Deep organized learning, or deep learning in its short form, has opened a new chapter in the field of machine learning since 2006 [5]. Deep learning, in its most general form, is a machine learning method that is used to solve problems and perform behavior such as analysis, inference, observation, and learning using large quantities of data, according to the descriptions in the literature. They can be in different hierarchical systems than conventional machine learning algorithms [9].

Artificial neural networks (ANNs) are a subfield of deep learning (ANN). It's a type of artificial neural network that uses nonlinear transformations to get a specific output value from raw data. The backpropagation algorithm is used in deep learning to investigate the dynamic structure of multidimensional data sets. This is accomplished by comparing the values of the parameters measured in each layer to the values obtained in the previous layer. Multi-layer neural networks have an input layer that represents the inputs, hidden layers that transform the information from the input layer into an output, and an output layer that converts the results from the last hidden layer into output values. The success rate in the fields of natural language processing, image processing, visual object identification, and drug discovery has increased significantly thanks to deep learning methods [10].

### 2.3. Data Analyses and Modelling

Qualitative data were expressed as number (percentage), while quantitative data were expressed as mean ± standard deviation, and median (minimum-maximum). The Kolmogorov-Smirnov test was used to assess conformity to the normal distribution. The Kruskal-Wallis, one-way analysis of variance, and Pearson chi-square tests were used to determine whether there is a statistically significant difference between the "Nondemented," "Demented," and "Converted" groups, which are the categories of the dependent/target variable (dementia) in terms of independent variables. Statistics were considered significant at  $p < 0.05$ . All analyses used the IBM SPSS Statistics 26.0 package program.

A 10-fold cross-validation procedure was employed to determine the model's validity. The 10-fold cross-validation method divides the entire set of data into 10 equal portions. The method is performed ten times using one portion as a test set and the other nine parts as training data sets. In addition, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score criteria were taken into account for the performance evaluation of the model.

## 3. RESULTS

Descriptive statistics for the independent variables examined in this study are given in Table 2. According to the findings in Table 2; there is a statistically significant difference between the dependent / target variable groups in terms of Mr delay, educ, mmse, cdr, nwbv, age variables ( $p < 0.05$ ).

According to the findings in Table 3; There is a statistically significant relationship between the mf, ses variables and the dependent / target variable (class) groups ( $p < 0.05$ ).

TABLE II  
DESCRIPTIVE STATISTICS FOR QUANTITATIVE INDEPENDENT VARIABLES

Variables	Grup						p-value
	Nondemented		Demented		Converted		
	Median (min-max)	Mean± Standard deviation	Median (min-max)	Mean± Standard deviation	Median (min-max)	Mean ± Standard deviation	
MR DELAY	631 <sup>a</sup> (0-2517)	-	491 <sup>b</sup> (0-2508)	-	706 (0-2639)	-	0.006**

<b>EDUC</b>	16 <sup>a</sup> (8-23)	-	13 <sup>b</sup> (6-20)	-	16 (12-20)	-	<0.001**
<b>MMSE</b>	29 <sup>a</sup> (26-30)	-	26 <sup>b</sup> (4-30)	-	29 (24-30)	-	<0.001**
<b>CDR</b>	0 <sup>ab</sup> (0-0.5)	-	0.5 <sup>b</sup> (0.5-2)	-	0.5 (0-0.5)	-	<0.001**
<b>ETIV</b>	1475 (1106-2004)	-	1477 (143-1957)	-	1423 (1264-1722)	-	0.517**
<b>NWBV</b>		0.741 <sup>ab</sup> ±0.038		0.716±0.032		0.724±0.035	<0.001***
<b>ASF</b>		1.191±0.144		1.197±0.0137		1.212±0.109	0.5683***
<b>Age</b>		77±8		76 <sup>b</sup> ±7		80±7	0.045***

\*: a: different according to the Demented group, b: different according to the Converted group  
 \*\*: Kruskal Wallis test, \*\*\*: One way analysis of variance test

TABLE III  
 DESCRIPTIVE STATISTICS FOR QUALITATIVE INDEPENDENT VARIABLES

Variables		Group			p- value**
		Nondemented	Demented	Converted	
		Count (%)	Count (%)	Count (%)	
Visit	1	72 (37.9)	64 (43.8)	14 (37.8)	0.339
	2	70 (36.8)	62 (42.5)	12 (32.4)	
	3	34 (17.9)	16 (11.0)	8 (21.6)	
	4	10 (5.3)	3 (2.1)	2 (5.4)	
	5	4 (2.1)	1 (0.7)	1 (2.7)	
MF	Male	61 (32.1)	86 (58.9)	13 (35.1)	<0.001
	Female	129 (67.9)	60 (41.1)	24 (64.9)	
SES	1	41 (21.6)	26 (20.5)	21 (56.8)	<0.001
	2	71 (37.4)	25 (19.7)	7 (18.9)	
	3	42 (22.1)	33 (26.0)	7 (18.9)	
	4	34 (17.9)	38 (29.9)	2 (5.4)	
	5	2 (1.1)	5 (3.9)	0 (0.0)	

\*\* : Pearson chi-square test

TABLE IV  
 DESCRIPTIVE STATISTICS FOR QUALITATIVE INDEPENDENT VARIABLES

Metrics	Tissue types	Nondemented	Demented	Converted
Accuracy		0.909	0.947	0.893
Sensitivity		0.908	0.946	0.824
Specificity		0.988	0.978	0.935
Positive predictive Value		0.989	0.966	0.378
Negative predictive Value		0.896	0.965	0.991
F1-score		0.947	0.956	0.519

\*\* : Pearson chi-square test

The highest metric values among the performance criteria achieved for group variable types were calculated for Demented; and we found that Accuracy, Sensitivity, Specificity, Positive predictive value, Negative predictive

Value, F1-score were 0.947, 0.946, 0.978, 0.966, 0.965 and 0.956 respectively.

The values showing the significance of the variables are shown in Table 5. The five most important variables were found cdr (0.1184), mmse (0.1005) and visit (0.0990), etiv (0.0909), ses (0.0903).

TABLE V  
 DESCRIPTIVE STATISTICS FOR QUALITATIVE INDEPENDENT VARIABLES

EXPLANATORY VARIABLES	Importance
CDR	0.1184
MMSE	0.1005
VISIT	0.0990
ETIV	0.0909
SES	0.0903
MRDELAY	0.0862
MF	0.0855
EDUC	0.0854
ASF	0.0846
NWBV	0.0807
Age	0.0779

#### 4. DISCUSSION

The rapid aging in the world and the increase in the elderly population as a result of this is a global problem. By 2050, it is estimated that the population aged 60 and over will be around 2 billion, and the population aged 80 and over will be around 400 million [11,12]. It is stated that with this situation, the incidence of dementia will increase sharply and this disorder will create great challenges for public health and aged care systems in all countries of the world. For this reason, it has become very important to determine the management and treatment procedures of dementia, to reduce the level of progression of the disease and to increase the quality of life of individuals exposed to the disease [13]. In this study, it was aimed to determine the risk factors associated with dementia by using the data set that included the data of patients with dementia, non-dementia, and those who developed dementia during follow-up.

Machine learning is a branch of science concerned with the design and creation of algorithms that enable computers to learn from various types of data. Machine learning is both a database problem and a branch of artificial intelligence that enables current data to be used to predict and simulate future events [14]. Deep learning is a form of machine learning that is becoming increasingly common. The Deep learning approach has grown in popularity as graphics processing units (GPU) have improved [9]. The deep learning approach is designed inspired by the working principle of the brain. It is a process that was developed using the example of human learning systems and consists of several hidden layers and neurons. The deep learning method is used to process large data sets by contributing hardware features that evolve over time (especially the participation of graphics processors in the calculation). Motion detection, face recognition, health technologies, object recognition, and object detection are all fields where deep learning is used [15].

In the study, deep learning method was applied to the Dementia data set which is an open-source data set. When three different types of group variable were classified, the accuracy value was obtained as 91.96. In a study using the same data set, the highest accuracy obtained from

classification models made with different machine learning methods and voting method is 91.17% [16]. In the current study, higher accuracy was obtained in the classification made with the deep learning model. Demented is the one with the best performance metrics from three different categories for the group variable. Deep learning method gave successful predictive results in the classification of dementia according to the results of the performance criteria calculated in this study. In addition, the importance of risk factors related to dementia was obtained with the experimental findings. The most important 3 factors that may be associated with having a dementia were obtained as cdr, mmse, visit.

Detection and classification of dementia have a very important place in the diagnosis of neurogenerative disorders. As a result, when the findings obtained from this study were examined, the dementia dataset, which consisted of imaging data and information about patients with clinical data, was classified with high accuracy using the deep learning method. The risk factors for dementia were determined with the variable importance values obtained as a result of the model.

## REFERENCES

- [1] G. Demir, "Demans Ve Hemşirelik Bakımı," Black Sea Journal Of Health Science, Vol. 1, Pp. 35-39.
- [2] M. Braun, U. Scholz, B. Bailey, S. Perren, R. Hornung, and M. Martin, "Dementia caregiving in spousal relationships: a dyadic perspective," Aging and mental health, vol. 13, pp. 426-436, 2009.
- [3] M. Prince, M. Guerchet, and M. Prina, "Alzheimer's Disease International (2013b) Policy Brief for Heads Of Government: The Global Impact of Dementia 2013-2050," Alzheimer's Disease International, London, 2014.
- [4] Ö. G. D. İ. Akyar, "Demanslı hasta bakımı ve bakım modelleri," Hacettepe Üniversitesi Hemşirelik Fakültesi Dergisi, vol. 18, pp. 79-88, 2011.
- [5] L. Deng and D. Yu, "Deep learning: methods and applications," Foundations and trends in signal processing, vol. 7, pp. 197-387, 2014.
- [6] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," Large-scale kernel machines, vol. 34, pp. 1-41, 2007.
- [7] A. Şeker, B. Diri, and H. H. Balık, "Derin öğrenme yöntemleri ve uygulamaları hakkında bir inceleme," Gazi Mühendislik Bilimleri Dergisi (GMBD), vol. 3, pp. 47-64, 2017.
- [8] K. Umut, A. YILMAZ, and Y. Dikmen, "Sağlık alanında kullanılan derin öğrenme yöntemleri," Avrupa Bilim ve Teknoloji Dergisi, pp. 792-808, 2019.
- [9] K. Kayaalp and A. Süzen, "Derin Öğrenme ve Türkiye'deki Uygulamaları," Yayın Yeri: IKSAD International Publishing House, Basım sayısı, vol. 1, 2018.
- [10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, pp. 436-444, 2015.
- [11] Y. Onadja, N. Atchessi, B. A. Soura, C. Rossier, and M.-V. Zunzunegui, "Gender differences in cognitive impairment and mobility disability in old age: a cross-sectional study in Ouagadougou, Burkina Faso," Archives of gerontology and geriatrics, vol. 57, pp. 311-318, 2013.
- [12] Ü. Görgülü and N. Akdemir, "İleri evre kanser hastalarına bakım verenlerin yorgunluk ve uyku kalitesinin değerlendirilmesi," Genel Tıp Dergisi, vol. 20, 2010.
- [13] M. N. Özata Değerli, "Orta Evre Alzheimer Tip Demans Tanılı Bireylerde Demansa Bağlı Davranışsal Ve Psikolojik Semptomlar İle Duyusal İşleme Süreci Arasındaki İlişkinin İncelenmesi," 2022.
- [14] E. Alpaydin, Introduction to machine learning: MIT press, 2020.
- [15] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural networks, vol. 61, pp. 85-117, 2015.
- [16] S. Bharati, P. Podder, D. N. H. Thanh, and V. Prasath, "Dementia classification using MR imaging and clinical data with voting based machine learning models," Multimedia Tools and Applications, pp. 1-22, 2022.

## BIOGRAPHIES

**Onural Özhan** obtained her BSc. degree in pharmacy from Inonu University in 2009. He received MSc. degree in pharmacognosy from Ankara University in 2012. He received Ph.D. degree in pharmacology from Inonu University in 2019. In 2015, he joined the Department of pharmacology at Inonu University as a research assistant. Currently, he continues to work as an asst. prof. at Inonu University. His research interests are the cardiovascular system, oxidative stress, and ischemia-reperfusion.

**Zeynep Küçükakçalı** obtained her BSc. degree in mathematics from Çukurova University in 2010. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.

**İpek Bahkçı Çiçek** obtained her BSc. degree in mathematics from Çukurova University in 2010. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.



# Prediction of Placenta Accreta Spectrum by Machine Learning Methods and Determination of Candidate Biomarkers

<sup>1</sup>Seyma Yasar , <sup>1</sup>Saim Yologlu 

<sup>1</sup>Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey. (e-mail: {seyma.yasar, saim.yologlu}(1)@inonu.edu.tr).

## ARTICLE INFO

**Received:** Sep.,09.2022

**Revised:** Oct.,19.2022

**Accepted:** Oct.,29.2022

### Keywords:

Placenta Accreta Spectrum  
Biomarker  
Machine learning  
XGBoost  
Proteomics

**Corresponding author:** Seyma Yasar

✉ [seyma.yasar@inonu.edu.tr](mailto:seyma.yasar@inonu.edu.tr)

☎ +90 507 639 71 21

**ISSN:** 2548-0650

**DOI:** <https://doi.org/10.52876/jcs.1180200>

## ABSTRACT

Placenta accreta spectrum (PAS) disorders; Abnormal adhesion of placental villi to the myometrium associated with endometrial trauma or dysplasia. Placenta previa and previous cesarean section operations are two major risk factors for PAS disorders. It is usually diagnosed by ultrasound examinations performed during pregnancy follow-up. After this diagnosis is made, a very careful and strict pregnancy follow-up should be done. If the diagnosis is made during pregnancy, the delivery should be done by cesarean section and the bleeding that the mother will experience should be stopped with an appropriate method. However, no protein candidate to be used in clinical diagnosis has been found so far. The aim of this study is to identify candidate biomarkers that can be used in the diagnosis and follow-up of PAS with machine learning methods.

In this study, proteomic data obtained from 26 women with and without PAS were used. After using the Lasso method as the variable selection method, machine learning models (XGBoost, Adaboost) were created with 5-fold cross-validation. Accuracy, Balanced accuracy, Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, F1-Score, MCC and G-mean metrics were used in the performance evaluation of the models created.

When the performance metrics of the two models are compared, the best result belongs to the XGBoost machine learning model. Therefore, the Accuracy, Balanced accuracy, Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, F1-Score, MCC, and G-mean performance criteria for the XGBoost model are 0.962, 0.950, 1.00, 0.90, 0.94, 1.00, 0.97, 0.92, and 0.97, respectively.

As a result, considering the experimental results, it can be said that the created machine learning model is quite successful in classifying PAS. In addition, it can be said that KDR and AMH proteins are candidate biomarkers that can be used in the diagnosis and follow-up of PAS according to the significance of the variables related to the model.

## 1. INTRODUCTION

PLACENTA accreta is defined as abnormal trophoblast invasion of part or all of the placenta into the uterine wall myometrium. The diagnoses of placenta accreta, placenta percreta and placenta accreta, formerly known as placenta attachment anomaly, are currently called placenta accreta spectrum (PAS). It is known to cause obstetric hemorrhages, which often require blood transfusions after delivery and even threaten life. The most common known risk factor is the history of previous cesarean section and the number of previous cesarean sections. The risk increases as the number of previous cesarean sections increases. In a systemic review, when patients with a history of 1 cesarean section were

compared with patients with 4 and 5, the incidence of PAS increased from 0.3 to 6.74 [1]. Other risk factors are; maternal age, multiparity, history of previous uterine surgery, history of curettage, history of Asherman syndrome and presence of placenta previa [2]. On the other hand, in clinical practice, up to 50% of pregnancies with PAS are not diagnosed before delivery, resulting in increased morbidity [3,4]. Therefore, a new and improved paradigm is urgently needed for early and accurate diagnosis. PAS antenatal diagnosis is very important. Early diagnosis of patients with PAS will reduce poor outcomes, as it will be appropriate to follow up in hospitals with a comprehensive multidisciplinary approach for problems that may occur in the follow-up of patients with PAS and bleeding that may occur during delivery.

Machine learning, as a sub-branch of artificial intelligence, is a method based on the use of computer-aided mathematical models to reveal the relationships between data and make them meaningful. Thanks to the mathematical methods used, analyzes can be made on a large number of instantly updated data stacks. It is possible to draw meaningful conclusions from the tested data using machine learning methods and to interpret these results on untested data. It allows the learning activity to be done by computers in the process of data processing. It gives computers the ability to predict future events with the results obtained from the analysis of past data [5]. Recently, machine learning methods have been used frequently in the diagnosis and treatment of diseases in the field of health.

The aim of this study is to determine the proteins that can be used in the diagnosis and treatment of PAS and that can be biomarker candidates by applying tree-based methods such as XGBoost and Adaboost, which are machine learning methods, on open access PAS data.

## 2. MATERIAL and METHODS

### 2.1. Dataset

The open-source dataset used in this study consists of 1305 proteins obtained after proteomic analysis of plasma samples taken before birth from 16 patients with PAS at 35.1 gestational week and 10 control subjects at 35.5 gestational weeks with similar gestational ages [6]. Descriptive statistics for the subjects that make up the data set are given in Table 1.

TABLE I  
DESCRIPTIVE FEATURES OF PLACENTA ACCRETA SPECTRUM CASES AND CONTROL CASES

Variable	Invasive placenta (n=16)	Control (n=10)
Maternal age, year <sup>a</sup>	34.1 (32.4-37.2)	30.8 (30.0-36.7)
Body mass index at delivery, kg/m <sup>2</sup> <sup>a</sup>	33.7 (26.5-43.0)	28.5 (27.6-31.1)
Previous cesarean deliveries, n (%)		
0	1 (6.3)	4 (40.0)
1	7 (43.8)	5 (50.0)
2	2 (12.5)	0 (0.0)
3	2 (12.5)	1 (10.0)
4	4 (25.0)	0 (0.0)
Previa in current pregnancy, n (%)		
Yes	13 (81.3)	4 (40.0)
No	2 (12.5)	5 (50.0)
Unknown	1 (6.3)	1 (10.0)
Gestational age at delivery, wka <sup>a</sup>	35.1 (34.6-35.4)	35.5 (35.2-35.7)

<sup>a</sup> Data are presented as median (interquartile range)

### 2.2. XGBoost

XGBoost algorithm; It was proposed by Tianqi Chen in 2016. It is an improved and higher performance version of the Gradient Boosting Machine algorithm based on decision trees. The XGBoost algorithm works 10 times faster than other algorithms and it does this thanks to its scalability. Its scalability is due to some algorithmic optimizations. XGBoost works much faster and with higher performance thanks to its parallel operation and hardware optimizations [7].

### 2.3. Adaboost

AdaBoost algorithm was developed by Robert Schapire and Yoav Freund in 1995. AdaBoost is a classifier method that can obtain a strong estimator as a result of combining the singular and weak estimators obtained in each iteration. Coefficients are assigned to the estimators generated at each iteration. In these steps, the weights of the weak estimators are increased. With the completion of the iterations, the estimators to be combined are obtained [8].

### 2.4. Performance Evaluation of the Models

The conformity of the variables in the data set to the normal distribution was evaluated with the Kolmogorov-Smirnov test. Quantitative variables are expressed as median (Inter Quantile Range), while qualitative variables are expressed as numbers (percentage). Lasso regression method, which is frequently used in large data sets, was used as a variable selection method because it is effective and fast [9]. Then, 5-fold cross validation method was used. Accuracy, Balanced accuracy, Sensitivity, Specificity, Positive predictive value, Negative predictive value, Matthews correlation coefficient (MCC), G-mean and F1-Score metrics in the performance evaluation of XGBoost and Adaboost machine learning models created to identify candidate biomarkers that can be used in the diagnosis and follow-up of PAS used.

## 3. RESULTS

As a result of the variable selection method applied to 1305 proteins in the open source data set used in the study, 125 proteins were included in the study. Classification matrices for XGBoost and Adaboost models created with these 125 proteins obtained to classify PAS are given in Table 2 and Table 3, respectively.

TABLE II  
CLASSIFICATION MATRIX OF THE TESTING STAGE FOR THE XGBOOST MODEL

		Real		
		Control	PAS	Total
Predicted	Control	9	0	9
	PAS	1	16	17
	Total	10	16	26

TABLE III  
CLASSIFICATION MATRIX OF THE TESTING STAGE FOR THE ADABOOST MODEL

		Real		
		Control	PAS	Total
Predicted	Control	9	2	11
	PAS	1	14	15
	Total	10	16	26

The performance metrics calculated through the obtained classification matrices are given in Table 4.

TABLE IV  
PERFORMANCE METRICS OF XGBOOST AND ADABOOST MODELS

METRICS	MACHINE LEARNING MODELS	
	XGBOOST	ADABOOST
	VALUE (95% CI)	VALUE (95% CI)
ACCURACY	0.96 (0.88-1)	0.89 (0.76-1.00)
BALANCED ACCURACY	0.95 (0.86-1)	0.88 (0.77-1.00)

SENSITIVITY	1.00 (0.79-1)	0.88 (0.62-0.98)
SPECIFICITY	0.90 (0.55-0.99)	0.90 (0.55-0.99)
POSITIVE PREDICTIVE VALUE	0.94 (0.71-0.99)	0.93 (0.68-0.99)
NEGATIVE PREDICTIVE VALUE	1.00 (0.66-1.00)	0.82 (0.48-0.98)
F1-SCORE	0.97 (0.90-1)	0.90 (0.79-1.00)
MCC	0.92 (0.82-1)	0.76 (0.60-0.93)
G-MEAN	0.97 (0.91-1)	0.87 (0.75-1.00)

Considering the performance metrics in Table 4, the values for the XGBoost model are higher. Therefore, the importance values of PAS-related proteins determined by this model are shown in Table 5 and Figure 1.

TABLE V  
THE IMPORTANCE VALUES OF PAS-RELATED PROTEINS DETERMINED BY XGBOOST MODEL

EXPLANATORY VARIABLES	IMPORTANCE VALUE	EXPLANATORY VARIABLES	IMPORTANCE VALUE
KDR	100	PPBP	21.356
AMH	65.246	TFF2	20.813
CNDP1	44.879	MAP2K4	20.079
CXCL8	40.416	FGA FGB FGG	19.0
TNFRSF11B	38.246	PPBP	18.477
IL18R1	31.612	FTCD	18.382
THBS1	31.362	SIRT2	18.345
CA1	24.294	HP	18.287
TFPI	22.534	TPT1	18.113
VEGFA	21.526	FGFR4	17.914

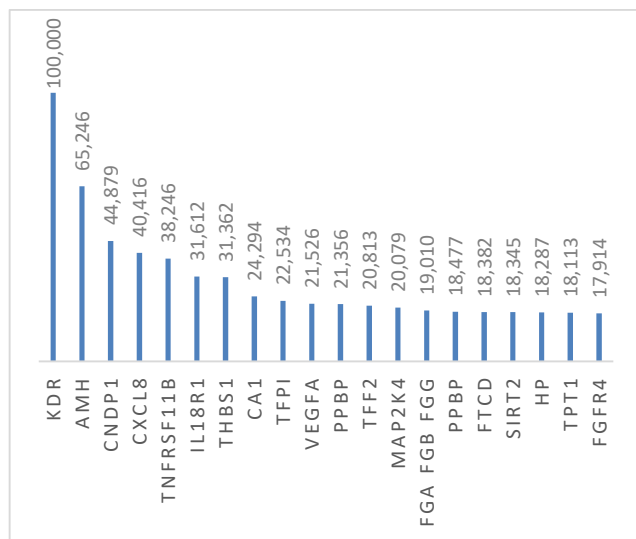


Fig. 2. The importance values for possible biomarkers

### 3. DISCUSSION

PAS is defined as one of the most serious disorders of pregnancy, as it is associated with a significant risk of massive obstetric hemorrhage and thus a high risk of admission to the maternal intensive care unit, reoperation, and prolonged hospitalization. On the other hand, early detection of the placental accreta spectrum (PAS), which is defined as the abnormal invasion of trophoblasts into the myometrial layer at different depths of invasion, is very important for the most

appropriate surgical management and to prevent bleeding during delivery [10].

In the case of PAS, the best acceptable approach is cesarean hysterectomy [11]. However, this is not accepted by most patients. PAS can be diagnosed with high sensitivity and specificity by ultrasonography in the second and third trimesters [12, 13]. On the other hand, MRI is reported to be beneficial in cases of posterior PAS. However, MRI is expensive and requires expertise not commonly found in the diagnosis of accreta. In two studies that directly compared the two imaging modalities, MR was not found to be superior to ultrasonography [14, 15]. Therefore, there is a need for easier and cheaper diagnostic methods to be used in clinical diagnosis.

Studies in which PAS and machine learning models are integrated are very limited in the literature. In a study involving 727 women with PAS, a machine learning model was created combining baseline and perioperative variables, and the highest AUC values at which the model was evaluated were found to be 0.90 [16]. In another study, the accuracy values of different machine learning algorithms (Random Forest, K-Nearest Neighbor, Naive Bayes, Multilayer Perceptron) created using region of interest (ROI) were obtained as 95.6, 98.1, 80.5 and 88.6, respectively [17].

In this study, machine learning methods (XGBoost, Adaboost) created using proteomic data from 26 subjects (PAS=16, Control=10) were used to classify PAS. According to the experimental results obtained, the accuracy values obtained from the XGBoost and Adaboost models are 0.962 and 0.885, respectively.

As a result, the proteomic dataset used in the study and the XGBoost model have a very high classification performance. Therefore, considering the variable significance obtained from the model result, it can be said that KDR, AMH proteins are candidate biomarkers that can be used in the diagnosis of PAS. In addition, as a result of the confirmation analyzes to be made, it is predicted that these two proteins can be used in the clinical diagnosis of PAS.

### REFERENCES

- [1] Marshall, N.E., Fu R., Guise, J.M.(2011). Impact of multiple cesarean deliveries on maternal morbidity: a systematic review. American journal of obstetrics and gynecology. 205(3):262. e1-. e8.
- [2] Miller, R., Timor-Tritsch, I.E., Gyamfi-Bannerman, C., Medicine SfM-F. (2020) Society for Maternal-Fetal Medicine (SMFM) consult series# 49: cesarean scar pregnancy. American Journal of Obstetrics and Gynecology. 222(5):B2-B14.
- [3] Bailit, J.L., Grobman, W., Rice, M.M., Reddy, U.M., Wapner, R.J., Varner, M.W., et al. (2015) Morbidly adherent placenta treatments and outcomes. Obstetrics and gynecology. 125(3):683.
- [4] Fitzpatrick, K., Sellers, S., Spark, P., Kurinczuk, J., Brocklehurst, P., Knight, M. (2015) The Management and Outcomes of Placenta Accreta, Increta, and Percreta in the United Kingdom: A Population-based Descriptive Study. Obstetric Anesthesia Digest. 35(1):24-5.
- [5] Greener, J.G., Kandathil, S.M., Moffat, L., Jones, D.T. (2022) A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology. 23(1):40-55.
- [6] Shainker, S.A., Silver, R.M., Modest, A.M., Hacker, M.R., Hecht, J.L., Salahuddin, S., et al. (2020) Placenta accreta spectrum: biomarker discovery using plasma proteomics. American Journal of Obstetrics and Gynecology. 223(3):433. e1-. e14.
- [7] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015) Xgboost: extreme gradient boosting. R package version 04-2. 1(4):1-4.
- [8] Schapire, R.E. (2013) Explaining adaboost. Empirical inference: Springer. p. 37-52.
- [9] Ranstam, J., Cook, J. (2018) LASSO regression. Journal of British Surgery. 105(10):1348.
- [10] Silver, R.M., Branch, D.W. (2018) Placenta accreta spectrum. New England Journal of Medicine. 378(16):1529-36.

- [11] Jauniaux, E., Alfirevic, Z., Bhide, A., Belfort, M., Burton, G., Collins, S., et al. (2018) Placenta Praevia and Placenta Accreta: Diagnosis and Management: Green-top Guideline No. 27a. *Bjog*. 126(1):e1-e48.
- [12] Berkley, E.M., Abuhamad, A.Z. (2013) Prenatal diagnosis of placenta accreta: is sonography all we need? *Journal of ultrasound in medicine*. 32(8):1345-50.
- [13] D'antonio, F., Iacovella, C., Palacios-Jaraquemada, J., Bruno, C., Manzoli, L., Bhide, A. (2014) Prenatal identification of invasive placentation using magnetic resonance imaging: systematic review and meta-analysis. *Ultrasound in Obstetrics & Gynecology*. 44(1):8-16.
- [14] Riteau, A-S., Tassin, M., Chambon, G., Le, Vaillant. C., de Laveaucoupet, J., Quere, M-P., et al. (2014) Accuracy of ultrasonography and magnetic resonance imaging in the diagnosis of placenta accreta. *PLoS One*. 9(4):e94866.
- [15] Balcacer, P., Pahade, J., Spektor, M., Staib, L., Copel, J.A., McCarthy, S. (2016) Magnetic resonance imaging and sonography in the diagnosis of placental invasion. *Journal of Ultrasound in Medicine*. 35(7):1445-56.
- [16] Shazly, S.A., Hortu, I., Shih, J-C., Melekoglu, R., Fan, S., Ahmed, FuA., et al. (2021) Prediction of clinical outcomes in women with placenta accreta spectrum using machine learning models: an international multicenter study. *The Journal of Maternal-Fetal & Neonatal Medicine*. 1-10.
- [17] Romeo, V., Ricciardi, C., Cuocolo, R., Stanzione, A., Verde, F., Sarno, L., et al. (2019) Machine learning analysis of MRI-derived texture features to predict placenta accreta spectrum in patients with placenta previa. *Magnetic resonance imaging*. 64:71-6.

**Şeyma YAŞAR** obtained her BSc. degree in mathematics from Gaziosman Paşa University in 2009. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning, proteomics, bioinformatics.

**Saim YOLOĞLU** was born in Malatya, Turkey. He received his BS in Statistics from Hacettepe University in 1980. He completed his master's degree from Hacettepe University Health Sciences Institute Biostatistics program in 1983 and his doctorate in 1988 in the same program. In the same year, he started to work as a lecturer in the Department of Biostatistics and Medical Informatics, Faculty of Medicine, İnönü University. Currently in the same department, Prof. Dr. works as.

## BIOGRAPHIES



# Genomic Biomarkers of Metastasis in Breast Cancer Patients: A Machine Learning Approach

Feyza Inceoğlu<sup>1</sup>, Fatma Hilal Yagin<sup>2</sup>

<sup>1</sup> Department of Biostatistics, Faculty of Medicine, Malatya Turgut Özal University, Malatya 44000, Turkey (e-mail: feyza.inceoglu@ozal.edu.tr).

<sup>2</sup> Department of Biostatistics, and Medical Informatics, Faculty of Medicine, Inonu University, Malatya 44000, Turkey (e-mail: hilal.yagin@inonu.edu.tr).

## ARTICLE INFO

Received: July.,03.2022

Revised: Sep., 21.2022

Accepted: Oct.,17.2022

### Keywords:

Breast Cancer,  
Gene Analysis,  
Machine Learning,  
XGBoost

Corresponding author: Feyza Inceoglu

✉ feyza.inceoglu@inonu.edu.tr

☎ +90 506 9316595

ISSN: 2548-0650

DOI: <https://doi.org/10.52876/jcs.1211185>

## ABSTRACT

One of the cancers with the highest incidence in the world is breast cancer (BC). The aim of this study is to identify candidate biomarker genes to predict the risk of distant metastases in patients with BC and to compare the performance of machine learning (ML) based models. In the study; Genomic dataset containing 24,481 gene expression levels of 97 patients with BC was analyzed. Biomarker candidate genes were determined by ML approaches and models were created with XGBoost, naive bayes (NB) and multilayer perceptron (MLP) algorithms. The accuracy values of XGBoost, NB and MLP algorithms were obtained as 0.990, 0.907 and 0.979, respectively. Our results showed that XGBoost has higher performance. The top five genes associated with BC metastasis were AL080059, Ubiquilin 1, CA9, PEX12, and CCN4. In conclusion, when the ML method and genomic technology are used together, the distant metastasis risk of patients with BC can be successfully predicted. The developed XGBoost model can distinguish patients with distant metastases. Identified biomarker candidate genes may contribute to diagnostic, therapeutic and drug development research in patients with metastases.

## 1. INTRODUCTION

THE cancer is the health problem with the highest incidence of cardiovascular diseases in the world. In the report published by the International Agency for Research on Cancer in 2012, with information on 184 countries, it was announced that 14.1 million new cancer cases and 8.2 million deaths were due to cancer [1, 2]. Breast cancer (BC) is a systemic disease that occurs as a result of the rapid spread and proliferation of cells in the mammary glands and ducts of the breast to tissues in different parts of the body. Cancer types with the highest incidence are lung (13%), breast (11.9%) and colon (9.7%) cancers in the world. The cancer types with the highest mortality are lung (19.4%), liver (9.1%) and stomach (8.8%) cancers. The predicted cancer cases for 2025 are 19.3 million [3, 4].

Risk factors affecting BC are demographic variables, hormonal system changes, lifestyles, and benign breast anomalies, environmental and hereditary factors [4, 5]. Gene analysis has an important place in the determination of genetic factors. Mutations in oncogene/antioncogene structures affect processes and formations. For BC; HER 2 and HER 1

(c-erbB-2 and 1), Ras, c-Mys, TP53, BRCA1, BRCA2, STK11, PTEN, CDH1, CHRK, ATM, PALB2 genes have been revealed in previous studies [5].

Gene expression, which forms the basis of analyzes in molecular structure studies for BC, is used in the diagnosis and treatment of BC [6, 7]. The use of large-scale genomic analyzes in today's studies reveals complex structures. In most of the studies, the relationship between tumor metastasis and widespread mutational structures were examined clonally [8]. Microarray technology developing depending on the developments in medical technologies; It offers researchers the opportunity to measure more than one gene structure at the same time. The gene structures of the diseased and non-disease groups are clearly analyzed for distinguishing features. Although it is difficult for researchers to analyze with multidimensional data, these difficulties disappear with ML [9].

ML offers researchers a dynamic analysis process and is frequently used. In ML, many different processes such as classification, summarization, clustering of data, methods of



establishing variable models are applied. ML also presents unobservable relationships in large databases to researchers. In the ML database system, this unobservable information is made with database technologies, modeling methods, statistical and mathematical analysis [10, 11].

When constructing ML classification models on high-dimensional microarray datasets, biomarker candidate genes associated with the disease of interest must first be identified. It will be possible to improve the performance metrics of the classification model to be created by identifying biomarker candidate genes [12].

Microarray technology, which is used together with ML, is a method that will facilitate the early diagnosis of BC. With Microarray technology, researchers can analyze the expression level of thousands of genes simultaneously and qualitatively. Accurate classification techniques for BC prognosis and treatment process with ML will also help clinicians. A prediction model is created by analyzing complex BC datasets with ML [13].

The aim of this study; to establish a supportive clinical prediction model for early diagnosis by identifying biomarker candidate genes that cause metastasis in patients with BC with ML approaches.

## 2. MATERIAL AND METHODS

### 2.1. Data

Gene expression data were obtained from the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) database and analyzed. The study examined 24,481 gene expression levels in 97 patients with BC, 46 (47%) of whom had developed distant metastases within 5 years and 51 (53%) were lymph node negative (pN0) who did not develop distant metastases [14, 15].

### 2.2. Methods

#### 2.2.1. Machine Learning Approach

Feature selection for genomic data are methods used in machine learning to shorten analysis time, identify disease/state of interest-associated biomarker candidate genes, and improve the performance of predictive models. In the study, recursive feature elimination (RFE) method based on logistic regression (LR) classifier was used to select candidate gene biomarkers associated with BC metastasis. RFE is a popular method as it is effective in selecting the features that are most relevant for estimating the target feature RFE is a wrapper feature selection method. The method selects features using a machine learning algorithm [16]. The LR used in the RFE is a method that uses the maximum likelihood estimation approach for regression and classification tasks and estimates the values of the parameters that maximize the probability obtained [17].

XGBoost, naïve bayes (NB), and multilayer perception (MLP) classifiers were used to predict BC metastasis after feature selection. An algorithm based on decision-tree (DT) and gradient-boosting (GB), XGBoost is a faster running algorithm compared to GB algorithms, with different regularization penalties to avoid overfitting [18]. NB is an algorithm based on conditional probability, which is assumed to be equal and independent from each other in the

classification of all attributes based on conditional probability [19].

MLP is a type of neural network used to support feed forward neural networks. In MLP, the input layer receives the signal to be processed and the output layer does the estimation and classification [20, 21].

The 10-fold cross validation (CV) method was used to validate the models. The k-fold CV splits the data into k blocks randomly and the algorithm uses the k-1 block as the training set and the remaining single block as the test set. The process continues until all blocks are used as a test set, and the average of all results represents the overall performance [22]. The performance of the models was evaluated with accuracy, Sensitivity, specificity, positive predictive value, negative predictive value, and F1-score, and the performance results of the models were compared.

## 3. RESULTS

In Table 1, the results of the performance measures for the models created for BC metastasis prediction are given. When Table 1 is examined, the accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values of the XGBoost model were obtained as 0.990, 0.978, 1.000, 1.000, 0.981, and 0.989, respectively. Accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1-score values for the NB model were obtained as 0.907, 0.978, 0.843, 0.849, 0.977 and 0.909, respectively.

In the MLP model, accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1-score values were obtained as 0.979, 0.957, 0.843, 0.849, 0.977 and 0.909, respectively. The results showed that the XGBoost model had higher performance in predicting BC metastasis compared to the NB and MLP models.

TABLE I  
The Performance of the Models

Metric	XGBoost	NB	MLP
Accuracy	0.990	0.907	0.979
Sensitivity	0.978	0.978	0.957
Specificity	1.000	0.843	1.000
Positive predictive value	1.000	0.849	1.000
Negative predictive value	0.981	0.977	0.962
F1-score	0.989	0.909	0.978

NB: Naive Bayes; MLP: Multilayer Perception

In Table 2 and Figure 1, the importance of genes according to their contribution to the prediction of BC metastasis of the XGBoost model was examined. According to the results of the study, the importance of AL080059, Ubiquilin 1, CA9, PEX12, and CCN4 genes were 100, 90.621, 53.731, 46.485, and 45.775, respectively.

TABLE II  
The Genes and Importance

Feature (Gene)	Importance
AL080059	100
Ubiquilin-1	90.621
CA9	53.731
PEX12	46.485
CCN4	45.775
NMU	40.069
SSX2	38.24
ALDH4A1	36.249
RAB5	35.927
ALDH6A1	34.281
ARL4D	31.954
PHF1	29.893
UBE2T	27.616
AF052087	24.889
KIAA0906	21.185
PRAME	20.766
TGFB3	19.256
CDKN3	8.622
SLC37A1	8.217
SCUBE2	7.583

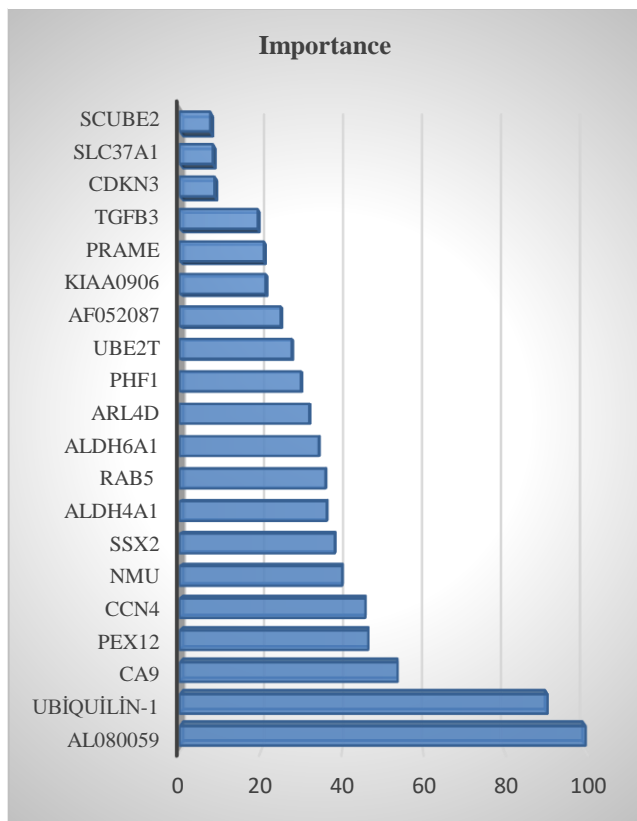


Fig. 1. Importance plot for biomarker genes

#### 4. DISCUSSION

Even though there have been significant advances in the treatment of BC in recent years, the prognosis for the majority of patients who have distant metastasis is still not good. Patients diagnosed with BC who are at the same stage of the disease may have very different responses to treatment and very different outcomes overall. In order to pave the way for earlier detection of metastasis and more effective treatments, it is critical to have an in-depth understanding of the molecular phenotype of distant metastasis. Based on this, the purpose of this study was to predict the presence of distant metastases in BC patients using ML methods that were based on genomic biomarkers [14]. From this point of view, the aim of this study is to predict the presence of distant metastases in BC patients using ML methods based on genomic information and data.

Genomic data including 24,481 gene expression levels of 97 patients with and without metastasis were used in the study. Genomic data containing thousands of gene expression levels belonging to a small number of patients in ML models require some preprocessing at the analysis stage due to their high dimensionality. Therefore, before creating ML models in the study, biomarker candidate genes were selected by LR-based RFE method. As a result of the analyses, 20 genes associated with BC distant metastasis were identified. Models based on XGBoost, NB and MLP algorithms were created with these biomarker candidate genes. Our results showed that XGBoost has higher performance compared to NB and MLP models. The accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values of the XGBoost model were obtained as 0.990, 0.978, 1.000, 1.000, 0.981, and 0.989, respectively. The first five genes that made a preliminary significant contribution to the prediction performance of the optimal model, XGBoost, and thus to differentiate BC distant metastasis, were AL080059, Ubiquilin 1, CA9, PEX12, and CCN4. The results of the trait significance analysis showed that the significance of the AL080059, Ubiquilin 1, CA9, PEX12, and CCN4 genes were 100, 90,621, 53,731, 46,485, and 45,775, respectively.

Our biomarker gene selection results were similar to the literature. In a study, it was reported that the AL080059 gene had a significant difference in BC patients compared to healthy controls and could be a candidate for a biomarker [23]. In different studies, it was found that UBQLN1 increased abnormally for BC [24, 25]. Similarly, in our study, it was reported that the Ubiquilin 1 (UBQLN1) gene showed a significant difference in BC patients and could be a biomarker candidate. CA9 has an important place in the distribution of tissues in the body and has been found to be an important factor for BC [26]. In our study, CA9 was found to be the third most important risk variable for BC. PEX 12 is an important risk factor in BC as in liver cancer [27-29]. In our study, PEX12 was found to be one of the most important risk variables for BC. CNN proteins enable the activation of signal transduction within the cell. CNN4 gene is effective on cancer because it is effective in cell migration and increases epithelial-mesenchymal transition [30], which was found to be an important risk factor for BC as well in our study. NMU, which affects the invasive capacity of cancer cells, has been identified as a risk factor for BC in different studies [31, 32]. In the factors examined in our study, NMU was found to be a risk factor for BC.

As a result, genes identified in the early diagnosis and treatment of BC distant metastasis can be examined and the XGBoost model can successfully differentiate metastases.

## 5. CONCLUSIONS

In conclusion, with the methodology combined with genomic technology and ml method, the risk of distant metastasis of patients with bc can be successfully predicted. identified biomarker candidate genes may contribute to diagnosis, treatment and drug development research in patients with metastasis. the developed XGBoost model can distinguish patients with distant metastases.

## REFERENCES

- [1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6), 394-424.
- [2] Cancer, I. A. F. R. o., & Organization, W. H. (2012). *Breast Cancer Estimated Incidence, Mortality and Prevalence Worldwide*. Globocan 2012: World Health Organization.
- [3] Sowunmi, A., Alabi, A., Fatiregun, O., Olatunji, T., Okoro, U. S., & Etti, A. F. D. (2018). Trend of cancer incidence in an oncology center in Nigeria: *West African Journal of Radiology*, 25(1), 52.
- [4] Momenimovahed, Z., & Salehiniya, H. (2019). Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer: Targets and Therapy*, 11, 151.
- [5] Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., . . . Zhu, H.-P. (2017). Risk factors and preventions of breast cancer: *International Journal of Biological Sciences*, 13(11), 1387.
- [6] Hollecsek, B., Stegmaier, C., Radosa, J. C., Solomayer, E.-F., & Brenner, H. (2019). Risk of loco-regional recurrence and distant metastases of patients with invasive breast cancer up to ten years after diagnosis—results from a registry-based study from Germany: *Bmc Cancer*, 19(1), 1-14.
- [7] Anwar, S. L., Avanti, W. S., Nugroho, A. C., Choridah, L., Dwianingsih, E. K., Harahap, W. A., . . . Wulaningsih, W. (2020). Risk factors of distant metastasis after surgery among different breast cancer subtypes: a hospital-based study in Indonesia: *World Journal of Surgical Oncology*, 18(1), 1-16.
- [8] Savas, P., Teo, Z. L., Lefevre, C., Flensburg, C., Caramia, F., Alsop, K., . . . Silva, M. J. (2016). The subclonal architecture of metastatic breast cancer: results from a prospective community-based rapid autopsy program “CASCADE”: *PLoS medicine*, 13(12), e1002204.
- [9] Xu, C., Meng, L. B., Duan, Y. C., Cheng, Y. J., Zhang, C. M., Zhou, X., & Huang, C. B. (2019). Screening and identification of biomarkers for systemic sclerosis via microarray technology: *International journal of molecular medicine*, 44(5), 1753-1770.
- [10] Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare: Paper presented at the Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics.
- [11] Yağın, F. H., Yağın, B., Arslan, A. K., & Çolak, C. (2021). Comparison of Performances of Associative Classification Methods for Cervical Cancer Prediction: Observational Study: *Turkiye Klinikleri Journal of Biostatistics*, 13(3).
- [12] Khaire, U. M., & Dhanalakshmi, R. (2020). High-dimensional microarray dataset classification using an improved adam optimizer (iAdam): *Journal of Ambient Intelligence and Humanized Computing*, 11(11), 5187-5204.
- [13] Vaka, A. R., Soni, B., & Reddy, S. (2020). Breast cancer detection by leveraging Machine Learning: *ICT Express*, 6(4), 320-324.
- [14] Akbulut, S., Yağın, F. H., & Çolak, C. (2022). Prediction of COVID-19 Based on Genomic Biomarkers of Metagenomic Next-Generation Sequencing (mNGS) Data using Artificial Intelligence Technology: *Erciyes Medical Journal*.
- [15] Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., . . . Witteveen, A. T. (2002). Gene expression profiling predicts clinical outcome of breast cancer: *nature*, 415(6871), 530-536.
- [16] Lee, M., Lee, J.-H., & Kim, D.-H. (2022). Gender recognition using optimal gait feature based on recursive feature elimination in normal walking: *Expert Systems with Applications*, 189, 116040.
- [17] Yilmaz, R., & Yağın, F. H. (2022). Early detection of coronary heart disease based on machine learning methods: *Medical Records*, 4(1), 1-6.
- [18] Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers: Environment and Urban Systems*, 96, 101845.
- [19] Paksoy, N., & Yagin, F. H. (2022). Artificial Intelligence-based Colon Cancer Prediction by Identifying Genomic Biomarkers: *Medical Records*, 4(2), 196-202.
- [20] Yilmaz, R., & Yagin, F. H. (2021). A comparative study for the prediction of heart attack risk and associated factors using MLP and RBF neural networks: *The Journal of Cognitive Systems*, 6(2), 51-54.
- [21] Akbulut, S., Yagin, F. H., & Colak, C. (2022). Prediction of Breast Cancer Distant Metastasis by Artificial Intelligence Methods from an Epidemiological Perspective: *Istanbul Medical Journal*, 23(3).
- [22] Perçin, İ., Yağın, F. H., Arslan, A. K., & Çolak, C. (2019). An interactive web tool for classification problems based on machine learning algorithms using java programming language: data classification software: Paper presented at the 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT).
- [23] Song, Q., Jing, H., Wu, H., Zou, B., Zhou, G., & Kambara, H. (2016). Comparative Gene Expression Analysis of Breast Cancer-Related Genes by Multiplex Pyrosequencing Coupled with Sequence Barcodes *Advances and Clinical Practice in Pyrosequencing*: Springer, 315-325.
- [24] Feng, X., Cao, A., Qin, T., Zhang, Q., Fan, S., Wang, B., . . . Li, L. (2021). Abnormally elevated ubiquitin-1 expression in breast cancer regulates metastasis and stemness via AKT signaling: *Oncology Reports*, 46(5), 1-14.
- [25] Jantrapimom, S., Lo Piccolo, L., Pruksakorn, D., Potikanond, S., & Nimlamool, W. (2020). Ubiquitin networking in cancers: *Cancers*, 12(6), 1586.
- [26] Hu, Z., Li, X., Yuan, R., Ring, B. Z., & Su, L. (2010). Three common TP53 polymorphisms in susceptibility to breast cancer, evidence from meta-analysis: *Breast cancer research and treatment*, 120(3), 705-714.
- [27] Moelans, C. B., De Weger, R. A., & Van Diest, P. J. (2010). Absence of chromosome 17 polysomy in breast cancer: analysis by CEP17 chromogenic in situ hybridization and multiplex ligation-dependent probe amplification: *Springer*, 120, 1-7.
- [28] Smeets, A., Daemen, A., Vanden Bempt, I., Gevaert, O., Claes, B., Wildiers, H., . . . De Moor, B. (2011). Prediction of lymph node involvement in breast cancer from primary tumor tissue using gene expression profiling and miRNAs: *Breast cancer research and treatment*, 129(3), 767-776.
- [29] Daskalaki, I., Gkikas, I., & Tavernarakis, N. (2018). Hypoxia and selective autophagy in cancer development and therapy: *Frontiers in Cell and Developmental Biology*, 6, 104.
- [30] Nivison, M. P., & Meier, K. E. (2018). The role of CCN4/WISP-1 in the cancerous phenotype: *Cancer Management and Research*, 10, 2893.
- [31] Wu, Y., McRoberts, K., Berr, S., Frierson, H., Conaway, M., & Theodorescu, D. (2007). Neuromedin U is regulated by the metastasis suppressor RhoGDI2 and is a novel promoter of tumor formation, lung metastasis and cancer cachexia: *Oncogene*, 26(5), 765-773.
- [32] Garczyk, S., Klotz, N., Szczepanski, S., Denecke, B., Antonopoulos, W., Von Stillfried, S., . . . Dahl, E. (2017). Oncogenic features of neuromedin U in breast cancer are associated with NMUR2 expression involving crosstalk with members of the WNT signaling pathway: *Oncotarget*, 8(22), 36246.

## BIOGRAPHIES

**Feyza İnceoğlu** obtained her BSc. degree in statistics from Gazi University (GU) in 2010. Feyza İnceoğlu obtained her BSc degree in statistics from Gazi University (GU) in 2010. She received the MSc ve PhD diploma in Biostatistics and Medical Informatics from the İnönü University in 2013 and 2018. She started to work as a statistician in the Republic of Turkey State Railways in 2012-2021. In 2021, she started to work as an Assistant Professor in Malatya Turgut Özal University, Faculty of Medicine, Department of Biostatistics and is still working. She works in the fields of validity, reliability, artificial intelligence.

**Fatma Hilal Yagin** obtained her BSc. degree in Statistics from Gazi University in 2017. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2020. She currently continues Ph.D. education in biostatistics and medical informatics from the Inonu University. In 2019, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning, and image processing.



# Framework for A Foreign Language Teaching Software for Children Utilizing AR, Voicebots and ChatGPT (Large Language Models)

<sup>1</sup>Oguzhan Topsakal , <sup>2</sup>Elif Topsakal 

<sup>1</sup>Florida Polytechnic University, Lakeland, Florida, USA. (e-mail: [otopsakal@floridapoly.edu](mailto:otopsakal@floridapoly.edu)).

<sup>2</sup>University of South Florida, Tampa, Florida, USA. (e-mail: [elift@usf.edu](mailto:elift@usf.edu)).

## ARTICLE INFO

**Received:** Dec., 04, 2022

**Revised :** Dec., 20, 2022

**Accepted:** Dec., 29, 2022

### Keywords:

Foreign Language Teaching

ChatGPT

Voicebot

Augmented Reality

Artificial Intelligence

Machine Learning

### Corresponding author:

Oguzhan Topsakal

✉ [otopsakal@floridapoly.edu](mailto:otopsakal@floridapoly.edu)

☎ +1

**ISSN:** 2548-0650

**DOI:** <https://doi.org/10.52876/jcs.1227392>

## ABSTRACT

The cognitive capabilities of children develop during the early years of their life. Preschool years are considered the best time for children to learn a second language. An increasing number of parents would like their children to start learning a foreign language at an early age. However, engaging little children with learning activities is challenging. In this study, we propose a framework for developing a foreign language learning software tool utilizing Augmented Reality (AR), Voicebots, and ChatGPT (an AI tool utilizing the Large Language Models) technologies to provide a unique product for young children to teach a foreign language. With AR and Voicebots, the product can grab attention, motivate, and provide an entertaining learning environment. The capabilities of ChatGPT can be utilized to efficiently prepare the content for the software tool. We show how to utilize the capabilities of ChatGPT to generate interactive dialogs that can be hosted at Google DialogFlow. We also present important design principles for the software, such as gamification, social interaction, and surprise rewards. We believe the framework and the design principles we propose in this study can be a blueprint for developing highly effective foreign language teaching software.

## 1. INTRODUCTION

THE cognitive capabilities of children develop during the early years of their life. Research shows that learning a foreign language helps develop cognitive skills. Preschool years are considered the best time for children to learn a second language or to acquire a second language which is defined as learning a second language after a first language is already established [1, 2, 3, 4, 5]. Moreover, learning a foreign language has become essential, and many parents would demand their children start learning a foreign language at an early age. However, engaging young children with learning activities or tasks are challenging due to their short attention span compared to older children. Parents who want their children to learn another language might hire a tutor. However, hiring a teacher would not be affordable for most families, and a teacher will not be available anytime needed. Children become technology savvy and start enjoying apps on mobile devices even during toddler ages. Augmented Reality apps especially grab the attention of young children with their 'magical' features as they pop up a 3D virtual model of a

child's actual surroundings. Besides, the curious minds of young children like to engage in conversations with voice assistants like Apple Siri, Google Assistant, Microsoft Cortana, or Amazon Alexa. Once Augmented Reality technology is combined with Voice Assistant technologies, a software tool can be realized for young children to play and learn a foreign language. A 3D avatar can pop up in a child's surroundings to teach him/her a foreign language while having conversations or playing games. This 3D Avatar will essentially be a voicebot to provide children with a fun and engaging learning experience. Moreover, the voicebot will be much more cost-effective for the parents and will be available anytime to teach the child a foreign language.

Voicebots (sometimes called conversational bots) are the product of recent technological advancements in AI, Machine Learning, Natural Language Processing, Voice Recognition, and Speech to Text technologies [6]. Voicebots are currently used mainly as a customer-facing application to answer product-related questions, order a product, or schedule an event. We believe voicebots can be utilized for education and have considerable potential for supporting education in many

fields, including foreign language teaching. When designed and implemented carefully, voicebots can be a game changer for learning a foreign language, not only for children but for adults as well.

One of the challenges of such software is content creation. The software needs to have dialogs ready for interactions with the user. The interactions could be not only in the foreign language but both in the native and the foreign language. The interactions could aim to teach a specific subject, for example, to teach animals, numbers, shapes, colors, body parts, jobs, etc. All these variations increase the cost and time needed to develop such software.

Recently, ChatGPT became available for public use, and was surprised its users with its capabilities. It can generate content such as short stories and dialogs based on the brief instructions given by the user [7]. It can even generate code when the problem is defined. ChatGPT made content and code creation amazingly easy. While ChatGPT might be available for public use for a limited time, we believe that similar capable large language AI models will be available soon for content creation.

In this study, we propose a framework for developing a language learning software tool utilizing Augmented Reality (AR), Voicebots, and ChatGPT (an AI utilizing the Large Language Model) technologies to provide a unique product for young children to teach a foreign language. With AR and Voicebots, the product will grab attention, motivate, and provide an entertaining learning environment. The capabilities of ChatGPT (or a similar large language model) will be utilized to prepare the content for the software tool efficiently.

## 2. DESIGN PRINCIPLES of a FOREIGN LANGUAGE TEACHING SOFTWARE SYSTEM

Developing software for teaching a foreign language to children can be most effectively achieved if we can decide on the design principles correctly. Based on the research on similar language learning tools, we propose that the software should have the following components and properties [8]:

- Gamification
- Social Interaction
- Surprise Rewarding
- Application of Cognitive, Communicative, and Audio-Lingual Methodologies

### 2.1. Gamification

Due to the positive effect of the game on learning, digital games, which were previously for entertainment purposes, have also started to be used in education. Recent studies have focused on how games can be used effectively as teaching materials and how they will affect learning. As a result, numerous applications were developed that support learning with gamification design in many areas of education. One of these areas is foreign language education.

Today, gamification has become an area that is constantly developing and gaining popularity. One of the approaches to gamification design was proposed by Werbach and Hunter [9]. Their gamification design model includes 1) Determining the objectives, 2) Determining the target behaviors and how to measure them, 3) Identifying the players (grouped according to their abilities, fears, future plans, and game type), 4) Preparing the activity cycles (user should be able to progress within the system and the difficulty level of the game should be increased with each step taken) 5) Incorporating the

entertainment element for all types of users. 6) Using appropriate tools to progress according to the purpose, cycle, and user characteristics [10].

In many studies on gamification, it is underlined that gamification makes learning more interesting and therefore creates an effective and qualified learning process. At the same time, it is emphasized that it supports cognitive, emotional, and social development by increasing motivation as it provides an active participation environment for individuals [4]. By adopting gamification in the design principles of the software, we aim to increase the interest and curiosity of preschool and elementary school children in language learning.

Andrew Stott and Carman Neustaedter have included some dynamics and concepts that are applied in the learning environment and emphasized that they make success permanent by taking part in game design in their literature review called "Analysis of Gamification in Education" [7]. These concepts include freedom to fail, rapid feedback, progression, and storytelling [10]. Gamification techniques are used to bring game players into a competitive environment that has feedback, score collection, badges, leaderboards, winning prizes, etc. items were found [4, 11, 12, 13].

### 2.2. Social Interaction

Interaction is the event that occurs mutually between people. Social interaction, on the other hand, leads to dialogue and communication. Playing a game with more than one person rather than a single person will have a more successful learning process if played interactively in a social environment [14, 15, 16].

### 2.3. Surprise Rewarding

Surprise rewards need to be added to games in order to increase the time children spend in the application and to spend more learning time. Many products apply the technique of surprise rewards to make their user excited about the product. Nir Eyal mentioned in his book 'Hooked: How to Build Habit-Forming Products', a product should include a Surprise/Variable Reward System as one of the techniques used to create a desire to use it over and over again. [17]

### 2.4. Application of Cognitive, Communicative, and Audio-Lingual Methodologies

The software should build new knowledge on previous learning by following a hierarchical order in teaching foreign words and sentences. It should also provide a learner-centered approach and enable children to control their own learning and move them to the next stages in line with their own learning pace. The voicebot will help users focus on first listening and speaking skills. These can be achieved by applying the cognitive method, the communicative method, and the audio-lingual methods.

In the cognitive method, old learning is associated with new learning to achieve a meaningful learning experience. Instead of memorization and mechanical learning, teaching is carried out with participatory learning and methods that support learning that can create active memory.

In the communicative method, teaching is student-centered and teaching materials, environment, and activities are produced specifically for students. Care is taken to produce these materials in a way that reflects the culture of the target language and is functional in daily life.

According to the audio-linguistic method, the natural learning of the language begins with acquiring first listening and then speaking, reading, and writing skills. For this reason, the method gives more importance to listening, understanding and speaking than other skills. Structures and new words in the target language are taught through dialogues [8].

### 3. COMPONENTS of the FOREIGN LANGUAGE TEACHING SOFTWARE SYSTEM

We propose that an engaging foreign language learning software for young children needs to incorporate three main components:

- Augmented Reality (AR.)
- Voicebots (Chatbots supported by speech-to-text and text-to-speech)
- Content creation utilizing AI tools such as ChatGPT

#### 3.1. Augmented Reality (AR.)

Augmented reality (AR) is a technology that overlaps computer-generated data (images, sounds, etc.) onto the user's view. It enhances the user's perception of the real world by adding layers of digital information to it [18].

AR can be experienced through a smartphone or tablet with a camera and a specialized app or through AR headsets or glasses. When the app is open, and the camera is pointed at a specific image or object, the app recognizes it and displays additional information or digital content on top of the user's view.

AR has many potential applications in many areas, including education, entertainment, and industry. For example, AR can be used to visualize complex concepts, create immersive gaming experiences, or provide real-time instructions and assistance in manufacturing or repair processes. Augmented Reality apps especially grab the attention of young children with their 'magical' features as they pop up a 3D virtual model of a child's real surroundings. AR can be an engaging and interactive way for children to learn a foreign language. It allows children to actively participate in the learning process and makes it feel more like a game than a traditional lesson. AR can provide children with a context for learning new words and phrases. For example, children can learn about the objects in their surroundings and can remember the words better as they are associated with real-world objects and concepts. AR can engage multiple senses, such as sight, and sound which can make the learning experience more immersive and memorable.

AR technology can be embedded into mobile apps utilizing one of the several AR frameworks such as ARCore, ARKit, or Vuforia. ARCore is a part of the Android development software development kit (SDK) developed by Google. ARKit is a part of the software development kit (SDK) developed by Apple for iOS devices. On the other hand, Vuforia is a cross-platform AR development framework that allows developers to create AR experiences for a wide range of devices, including smartphones, tablets, and AR glasses. Vuforia can be utilized for developing AR applications for both Android and iOS platforms. For our prototype, we wanted to cover both iOS and Android devices, however did not want to develop two separate mobile apps via two separate code bases. Therefore, we decided to utilize Unity3D to develop a mobile app for AR using a single code base and then export it for iOS and Android. Unity is a cross-platform game development platform that allows developers to create 2D,

3D, AR, and VR games and experiences for a variety of platforms, including iOS, Android, Windows, and console platforms. Inside the Unity development environment, we utilized Vuforia software libraries for AR development and were able to export the Unity project as iOS and Android projects.

Several AR products have been introduced to teach children language concepts. These products include:

- Magical Animals [1],
- Neobear Popup Zoo ([www.neobear.com](http://www.neobear.com)),
- ZooKazam ([www.zookazam.com](http://www.zookazam.com)),
- Shifu Safari
- ([www.playshifu.com](http://www.playshifu.com)),
- Animal 4D+ ([www.octagonstudio.com](http://www.octagonstudio.com))

Only two of these focused on foreign language teaching and only one incorporated voice, text, and voice recognition in their product.

#### 3.2. Voicebots

The developments in machine learning, speech recognition, and text-to-speech technologies in recent years have made possible the developments in virtual personal assistants. For example, virtual personal assistants such as Apple Siri, Microsoft Cortana, Google Assistant, and Amazon Alexa, which have become widespread in use, provide many conveniences to people in their daily lives and have become routine helpers for operations such as sending e-mails, making phone calls, changing music, and sending messages. Conversational robots (chatbots) developed with similar technologies can answer the phone and perform special tasks such as taking orders and giving directions. Many commercial products with varying capabilities that help the development of chatbots emerged. Some of them are listed here:

- Dialogflow (<https://dialogflow.com>)
- IBM Watson Assistant (<https://www.ibm.com/cloud/watson-assistant/>)
- Amazon Lex (<https://aws.amazon.com/lex/>)
- ManyChat (<https://manychat.com/>)
- Chatfuel (<https://chatfuel.com/>)
- Wit.ai (<https://wit.ai/>)
- MindMeld (<https://www.mindmeld.com/>)
- Chatbot (<https://www.chatbot.com>)
- Azure Bot Service (<https://azure.microsoft.com/en-us/services/bot-service/>)

The listed products are not for foreign language learning but for producing chatbot applications. They take a text as input and output a text. Most of them lack voice-to-text or text-to-speech capabilities. Our goal is to develop voicebots. The only product on the market that utilizes voicebots to teach a foreign language is Mondly's Conversational Chatbot. However, Mondly's product is developed for adults and does not include any gamification features. It is more like a regular chatbot with speech recognition capabilities.

The chatbot development products listed above can be utilized to create voicebots once voice-to-text or text-to-speech functionality is integrated. In our prototype, we utilized DialogFlow (a product offered by Google) due to its ease of integration with the voice-to-text or text-to-speech functionality. Google's DialogFlow is also built on top of the

Google Cloud Platform, so it will be easy to scale when the number of users increases. It also supports many languages. There are four important concepts when developing a chatbot with Dialogflow:

- **Agent:** A Dialogflow agent is similar to a human call center agent. You train both to handle expected speaking scenarios, and your training doesn't have to be overly explicit. Agents also act as a high-level container for settings and data
- **Intents:** Links between what the user says and how you want the system to respond
- **Entities:** These are the subject categories your bot understands. For example, if your chatbot is for a clothing store, an entity category might be a product type: shoes, dresses, pants, hats, and another entity might be size: large, xl, small, medium, and another color.
- **Context:** Information obtained by the bot during the conversation. For example, the user's name or the users' favorite shoe color.

The interaction from the user is met by the agent and answered by the most appropriate intent (intent matching) defined for the agent. In the selection of intent, the interacting words and the previous words spoken during the conversation play an important role. With training phrases and parameters defined for 'Intent', it prepares the response (response) and delivers it to the user.

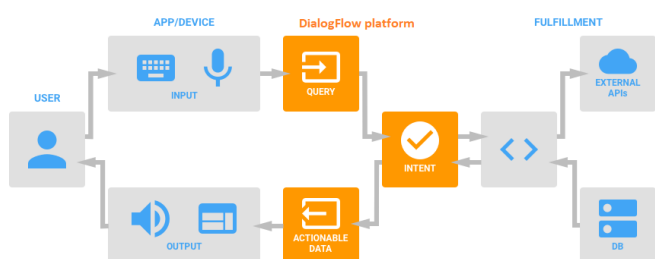


Fig.1. DialogFlow Interactions (Image Credit Google DialogFlow Documentation)

Figure 1 shows the flow of interactions between a user and a chatbot implemented using DialogFlow. Voice inputs from the user are translated into text and transmitted to the Dialogflow agent as input. The 'intent' corresponding to the text transmitted to the Dialogflow agent is selected based on the training phrases entered for intents. After the intent that will respond to the interaction from the user is determined (intent matching), the 'fulfillment' code, whose answer is determined for the intent, is prepared. This fulfillment code can be an external API and prepare the response after interacting with an external database.

Figure 2 shows the Interactions between the components of a voicebot. In response to the child's interactions (via speaking), the voice recognition software, embedded or called from within the software that is developed using the Unity development environment, translates the voice into text and transmits it to an agent in the Dialogflow. One of the many intents defined in this agent is selected via intent-matching algorithms to answer the incoming text. If a dynamic response is not required, the response is prepared by the intent and returned to the software without going to external APIs. If a dynamic response from the external environment is required, these external services are accessed via web interfaces and the prepared response is returned to the intent and then forwarded to the software.

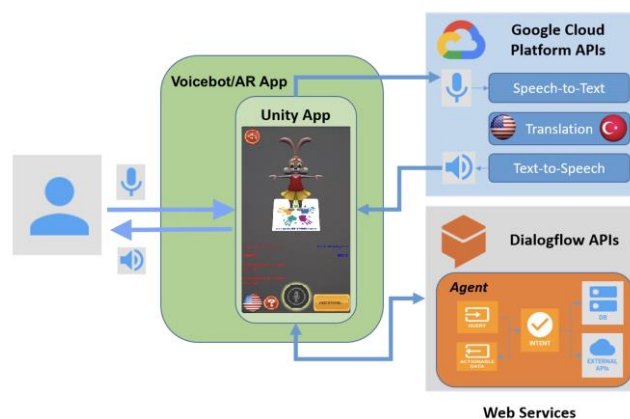


Fig.2. Interactions between the components of a voicebot

During the development process of Voicebot, Google Dialogflow product and Google Cloud Platform's speech-to-text, translation, and text-to-speech products were used. As seen in Figure 2, the user interacts with the mobile application developed in the Unity3D environment. After the user's voice is translated into text with 'Speech-to-Text', it is sent to DialogFlow. The answer text prepared in Dialogflow is converted into sound with 'Text-to-Speech' and transmitted to the user. In this interaction, the APIs of the Google Translation library are used when text is needed to be translated between languages.

### 3.3. Content Creation Utilizing AI tools (ChatGPT)

Designing conversations is a very important, challenging, and time-consuming task. The success of a voicebot heavily depends on how the interactions take place. To overcome the task of content generation, we propose to utilize an artificial intelligence (AI) tool, specifically, a large language model ChatGPT [19].

Large language models are trained to generate human-like text. They are called "large" because they are typically trained on very large datasets of text, often consisting of billions of words or more. The size of the data used to train a large language model can vary widely, but it is typically measured in billions of words. For example, the GPT-3 language model developed by OpenAI was trained on a dataset of over 8 billion words [20, 21], while the BERT language model [22], developed by Google was trained on a dataset of over 3 billion words.

Large language models have many potential applications, including natural language processing tasks such as translation, summarization, and question answering, as well as generating original content for a variety of purposes. They have been used to generate news articles, stories, and even poetry, and have the potential to revolutionize many fields that rely on text generation and understanding.

GPT and BERT are two of the well-known large language models. GPT (Generative Pre-training Transformer) is a transformer-based language model developed by OpenAI in 2018 that has been trained on a large dataset of web pages and books. BERT (Bidirectional Encoder Representations from Transformers) is another transformer-based language model developed by Google that has been trained on a large dataset of books and articles. New versions and variations of GPT and BERT have been introduced such as GPT-2, GPT-3, RoBERTa (Robustly Optimized BERT), and ALBERT (A Lite BERT) [21, 22, 23, 24].

Large language models have the potential to revolutionize the way that languages are taught and learned, by providing access to high-quality, customized, and personalized language learning materials.

We have tested ChatGPT to generate dialogues that would help in learning a foreign language. Some of the initial dialogues and subjects that are covered while teaching a foreign language, especially with young children, are introducing oneself, and conversations about animals, foods, vehicles, family members, body parts, vehicles, occupations, etc. We have asked the following prompts to ChatGPT, a large language AI model.

TABLE I<sub>SEP</sub>  
UTILIZING CHATGPT TO GENERATE DIALOGUES TO TEACH FOREIGN LANGUAGE

<b>ChatGPT Prompt:</b> <i>Generate a dialogue to teach English to young children whose native language is Turkish. Write Turkish meanings next to English sentences</i>
<b>ChatGPT Output:</b> <i>Teacher:</i> Hello, how are you? (Merhaba, nasılsın?) <i>Student:</i> I'm good, thank you. How are you? (İyiyim, teşekkür ederim. Sen nasılsın?) <i>Teacher:</i> I'm good too. What's your name? (Ben de iyiyim. Adın ne?) <i>Student:</i> My name is Selin. What's your name? (Adım Selin. Senin adın ne?) <i>Teacher:</i> My name is Mrs. Smith. Where are you from? (Benim adım Bayan Smith. Nereden geliyorsun?) <i>Student:</i> I'm from Turkey. And you? (Türkiye'den geliyorum. Sen nereden?) <i>Teacher:</i> I'm from the United States. Do you like learning English? (Amerika Birleşik Devletleri'nden geliyorum. İngilizce öğrenmeyi seviyor musun?) <i>Student:</i> Yes, I really like it. What do you want us to do today? (Evet, gerçekten seviyorum. Bugün ne yapmak istersin?) <i>Teacher:</i> Today we are going to learn some new words and phrases in English. Are you ready? (Bugün İngilizce'de yeni kelime ve cümleler öğreneceğiz. Hazır mısın?)

TABLE II<sub>SEP</sub>  
UTILIZING CHATGPT TO GENERATE DIALOGUES TO TEACH ANIMAL NAMES TO YOUNG CHILDREN

<b>ChatGPT Prompt:</b> <i>Using a similar style, generate a dialogue to teach English animal names to a Turkish young child.</i>
<b>ChatGPT Response:</b> <i>Teacher:</i> Look at this picture. Can you see the cat? (Bu resme bak. Kediyi görüyor musun?) <i>Student:</i> Yes, I see the cat. (Evet, kediyi görüyorum.) <i>Teacher:</i> Can you say "cat" in English? (İngilizce'de "kedi" diyebiliyor musun?) <i>Student:</i> Cat. (Kedi.) <i>Teacher:</i> Very good! Now, can you see the dog? (Çok iyi! Şimdi, köpeği görüyor musun?) <i>Student:</i> Yes, I see the dog. (Evet, köpeği görüyorum.) <i>Teacher:</i> Can you say "dog" in English? (İngilizce'de "köpek" diyebiliyor musun?) <i>Student:</i> Dog. (Köpek.) <i>Teacher:</i> Excellent! Let's learn some more animal names. Can you see the elephant? (Harika! Daha fazla hayvan ismini öğrenelim. Fil görüyor musun?) <i>Student:</i> Yes, I see the elephant. (Evet, fili görüyorum.) <i>Teacher:</i> Can you say "elephant" in English? (İngilizce'de "fil" diyebiliyor musun?) <i>Student:</i> Elephant. (Fil.) <i>Teacher:</i> Great job!

ChatGPT can also be used for generating code. In the framework, we proposed to utilize Chatbot creation products such as DialogFlow. We wanted to see if ChatGPT can put these dialogs into a format that can be easily imported to DialogFlow for usage. Therefore, we asked ChatGPT to create the one of the dialogues above in a format that can be imported into DialogFlow. ChatGPT was able to generate it

successfully. We present an excerpt of the response in Table 2 and 3.

TABLE III<sub>SEP</sub>  
AN EXCERPT OF THE INTERACTION WITH CHATGPT TO PREPARE THE CONVERSATION ABOUT FOREIGN LANGUAGE TEACHING IN A FORMAT THAT CAN READILY BE IMPORTED INTO DIALOGFLOW

<b>ChatGPT Prompt:</b> Can you generate the json file to import this dialog to Google DialogFlow?
<b>ChatGPT Response:</b> <pre>{   "id": "9c80f039-1b3d-46d3-b7f2-a3b3c3b75d71",   "name": "Animal Names",   "contexts": [],   "responses": [     {       "resetContexts": false,       "parameters": [],       "messages": [         {           "type": 0,           "lang": "en",           "speech": ["Look at this picture. Can you see the cat?"]         },         {           "type": 0,           "lang": "tr",           "speech": ["Bu resme bak. Kediyi görüyor musun?"]         }       ]     },     {       "resetContexts": false,       "parameters": [],       "messages": [         {           "type": 0,           "lang": "en",           "speech": ["Yes, I see the cat."]         },         ...       ]     }   ] }</pre>

#### 4. DISCUSSION

In recent years, studies have been carried out on teaching foreign languages with social robots. The 3D Avatars that will be utilized in the software along with augmented reality are similar to a robot. These studies revealed that the utilization of social robots improved the outcomes of foreign language learning tasks. It has been observed that learning achievements increase after spending more time with the robot [24]. According to these studies, the interactions with social robots (virtual 3D avatars) must be designed in a way that is pleasing, pedagogically sound, and challenging the limits of the learner [25, 26, 27, 28]. According to the studies, it would be meaningful to choose well-known word groups learned in the mother tongue in the selection of foreign language words that are aimed to be taught. The meaningful and interesting content of the teaching will also increase learning. In addition, it was recommended that the learning period be 10-15 weeks in 15-20 minutes doses every day and that a maximum of 6 words should be taught each time and each word should be repeated at least 10 times [28].

Software developed using the framework can scale in several dimensions including offering products in new countries with new native language options, offering products to teach new foreign languages, and offering new products to teach more content and concepts to cover more of the preschool curriculum for foreign language. Such a foreign language software will serve millions of families since there are approximately 26 million children between the ages of 4 and 8 in the Europe Union and there are about 17 million children between the ages of 4 and 8 in the USA. In



developing countries, learning a foreign language is considered even more essential.

## 5. CONCLUSION

We hope the framework we present here for developing foreign language teaching software will help bring innovative products to the market and will have a positive social impact. When children excel in a foreign language at an early age that means they communicate and understand each other better and possibly embrace the similarities and differences. Recent advancements in technology enable developers, researchers, and entrepreneurs to realize the components of the proposed framework more rapidly than ever. The cognitive development of young children can be increased by introducing them to foreign languages in a way that will grab their attention and engage them by utilizing Augmented Reality and Voicebots. The content for the foreign language teaching software can be conveniently produced using large language models such as ChatGPT.

## REFERENCES

- [1] Topsakal, Elif & Topsakal, Oguzhan. (2019). "Augmented Reality to Engage Preschool Children in Foreign Language Learning, Augmented Reality", *Virtual Reality, and Computer Graphics*, 286-294.
- [2] Kol, S. (2013). Erken Çocuklukta Bilişsel Gelişim ve Dil Gelişimi. *Sakarya Üniversitesi Eğitim Fakültesi Dergisi*, 21 (21).
- [3] Aydın, T. (2014). Dil Öğretimi ve Oyun - Çoklu Zekâ Teorisini Işığında. *Dinbilimleri Akademik Araştırma Dergisi*, 14 (1), 71-83.
- [4] Memiş, Muhammet. (2013). Yabancı Dil Öğretiminde Kullanılan Yöntemler, Kullanım Özellikleri ve Eleştiriler. *Turkish Studies*. 8. 297-318
- [5] Krashen, S. (1981) *Second Language Acquisition and Second Language Learning*. Pergamon Press, New York.
- [6] Valtolina, S., Barricelli, B.R., Gaetano, S.D., & Diliberto, P. (2018). Chatbots and Conversational Interfaces: Three Domains of Use. *CoPDA@AVI*.
- [7] Stott, A., & Neustaedter, C. (2013). Analysis of gamification in education. *Surrey, BC, Canada*, 8, 36.
- [8] Plass, J. L. (1998). Design and evaluation of the user interface of foreign language multimedia software: A cognitive approach. *Language Learning & Technology*, 2(1), 40–53. <http://dx.doi.org/10125/25031>
- [9] Werbach, Kevin, and Dan Hunter. (2015) *The gamification toolkit: dynamics, mechanics, and components for the win*. University of Pennsylvania Press.
- [10] Çağlar Özhan, Şeyma & Arkün Kocadere, Selay. (2015). Çevrimiçi öğrenme ortamlarında oyunlaştırma. *Eğitim Bilimleri ve Uygulama*. 14. 83-102
- [11] Dicheva, Darina & Dichev, Christo & Agre, Gennady & Angelova, Galia. (2015). Gamification in Education: A Systematic Mapping Study. *Educational Technology & Society*. 18. 75-88
- [12] Özkan, Zeynep & Samur, Yavuz. (2017). Oyunlaştırma Yönteminin Öğrencilerin Motivasyonları Üzerine Etkisi. *Ege Eğitim Dergisi*. 857-886.
- [13] Yıldırım, İ., & Demir, S. (2014). Gamification and education - Oyunlaştırma ve eğitim. *Journal of Human Sciences*, 11(1), 655-670.
- [14] Bayrakçı, M. (2013). Social Learning Theory and Its Educational Applications. *Sakarya Üniversitesi Eğitim Fakültesi Dergisi*, 0 (14), 198-210.
- [15] Karakaya, Z. (2011). Rol Oyunlarında Sosyal Etkileşim ve Dilsel Gelişim. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, 31 (2), 419-438
- [16] Pınar, Y. (2016). Anaokulunda Sosyal Etkileşim, Duygu ve Dil – İkinci Dil edinimi Ekseninde Bir Araştırma. *Eğitimde Nitel Araştırmalar Dergisi- Journal of Qualitative Research in Education*, 4(2), 29-51.
- [17] Eyal, Nir (2014), *Hooked: How to Build Habit-Forming Product*, USA
- [18] van Krevelen, D., & Poelman, R. (2010). A Survey of Augmented Reality Technologies, Applications and Limitations. *International Journal of Virtual Reality*, 9(2), 1–20. <https://doi.org/10.20870/IJVR.2010.9.2.2767>
- [19] ChatGPT, URL: <https://chat.openai.com/> Accessed Dec, 31<sup>st</sup>, 2022
- [20] Dale, Robert. (2021). GPT-3: What's it good for?. *Natural Language Engineering*. 27. 113-118. 10.1017/S1351324920000601.
- [21] Floridi, L. & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30:681–694 <https://doi.org/10.1007/s11023-020-09548-127>.
- [22] Klein, T., & Nabi, M. (2019). Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds. *ArXiv*, abs/1911.02365.
- [23] Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a markov random field language model. *CoRR*, abs/1902.04094
- [24] Rintjema, Emmy & van den Berghe, Rianne & Kessels, Anne & de Wit, Jan & Vogt, Paul. (2018). A Robot Teaching Young Children a Second Language: The Effect of Multiple Interactions on Engagement and Performance. 219-220.
- [25] Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., Montanier, J.-M., et al. (2019). Second language tutoring using social robots : a large-scale study. *HRI '19: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 497–505).
- [26] Van den Berghe, Rianne & Verhagen, Josje & Oudgenoeg-Paz, Ora & Van der Ven, Sanne & Leseman, Paul. (2018). Social Robots for Language Learning: A Review. *Review of Educational Research*
- [27] Verhoeven, Gijs & Catala, Alejandro & Theune, Mariet. (2019). Designing a Playful Robot Application for Second Language Learning. *Interactivity, Game Creation, Design, Learning, and Innovation*, 7th EAI International Conference, ArtsIT 2018, and 3rd EAI International Conference, DLI 2018, ICTCC 2018, Braga, Portugal, October 24–26, 2018, Proceedings
- [28] Belpaeme T, Vogt P, van den Berghe R, Bergmann K, Göksun T, de Haas M, Kanero J, Kennedy J, Küntay AC, Oudgenoeg-Paz O, Papadopoulos F, Schodde T, Verhagen J, Wallbridge CD, Willemsen B, de Wit J, Geçkin V, Hoffmann L, Kopp S, Kraemer E, Mamus E, Montanier JM, Oranç C, Pandey AK. (2018) Guidelines for Designing Social Robots as Second Language Tutors. *Int J Soc Robot*. 2018;10(3):325-341. doi: 10.1007/s12369-018-0467-6. Epub 2018 Jan 25. PMID: 30996752; PMCID: PMC6438435.

## BIOGRAPHIES

**Oguzhan TOPSAKAL** received his BS in Computer Engineering in 1996 from Istanbul Technical University, Turkey. He received his MS and Ph.D. in Computer Science from the University of Florida in 2003 and 2007. After gaining extensive experience in the software industry, Dr. Topsakal is currently an assistant professor in the Computer Science department at the Florida Polytechnic University since 2018. He teaches courses related to Machine Learning, Algorithm Design, Databases, and Mobile Development. Dr. Topsakal's research interests include applications of machine learning and deep learning in the medical field.

**Elif TOPSAKAL** got her bachelor's of education in science education. She got her master's degree in Science Education from the University of Florida in 2012. She continued her Ph.D. studies at Uludag University in Turkey in 2015 and completed all but her dissertation. Then she started the Ph.D. program in Educational Research and Measurement, at the University of South Florida in 2019. She is currently a Ph.D. Candidate and a Graduate Assistant at the College of Education at the University of South Florida. Her research interests include applications of machine learning and deep learning in the education field.



## PUBLICATION ETHICS

All who participate in producing The Journal of Cognitive Systems conduct themselves as authors, reviewers, editors, and publishers in accord with the highest level of professional ethics and standards. Plagiarism or self-plagiarism constitutes unethical scientific behaviour, and is never acceptable. By submitting a manuscript to this journal, each author explicitly confirms that the manuscript meets the highest ethical standards for authors and co-authors. **The undersigned hereby assign(s) to The Journal of Cognitive Systems (JCS) copyright ownership in the above paper, effective if and when the paper is accepted for publication by JCS, and to the extent transferable under applicable national law. This assignment gives JCS the right to register copyright to the paper in its name as claimant, and to publish the paper via any print or electronic medium.**

Authors, or their employers, in the case of works made for hire, retain the following rights.

- + all proprietary rights other than copyright, including patent rights
- + the right to make and distribute copies of the Paper for internal purposes
- + the right to use the material for lecture or classroom purposes
- + the right to prepare derivative publications based on the Paper, including books or book chapters, journal papers, and magazine articles, provided that publication of a derivative work occurs subsequent to the official date of publication by JCS.
- + the right to post an author-prepared version or an official version (preferred version) of the published paper on an internal or external server controlled exclusively by the author/employer, provided that (a) such posting is non-commercial in nature, and the paper is made available to users without charge; (b) a copyright notice and full citation appear with the paper, and (c) a link to JCS's official online version of the abstract is provided using the Document Object Identifier (DOI) link



# THE JOURNAL OF COGNITIVE SYSTEMS

an international, peer-reviewed, indexed, and  
open-access periodical

VOLUME 07, NUMBER 02

DECEMBER 2022

## CONTENTS

<b>O. Ozhan, and F.H.Yagin:</b> Machine Learning Approach for Classification of Prostate Cancer Based on Clinical Biomarkers,.....	17-20
<b>O. Ozhan, and Z. Kucukakcali :</b> Risk Prediction Model for Dementia by Deep Learning Using Clinical Data, .....	21-24
<b>S. Yasar, and S. Yologlu :</b> Prediction of Placenta Accreta Spectrum by Machine Learning Methods and Determination of Candidate Biomarkers, .....	25-28
<b>F. Inceoglu, and F.H. Yagin :</b> Genomic Biomarkers of Metastasis in Breast Cancer Patients: A Machine Learning Approach,.....	29-32
<b>O. Topsakal, and E. Topsakal:</b> Framework for A Foreign Language Teaching Software for Children Utilizing AR, Voicebots and ChatGPT (Large Language Models) , .....	33-38

