HARRAN ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ
AKADEMİK BİLİM DERGİSİ

HARRAN ÜNİVERSİTESİ MÜHENDİSLİK DERGİSİ

HARRAN UNIVERSITY JOURNAL OF ENGINEERING

e-ISSN: 2528-8733

HARRAN ÜNİVERSİTESİ MÜHENDİSLİK DERGİSİ

HARRAN UNIVERSITY JOURNAL of ENGINEERING

e-ISSN: 2528-8733 (ONLINE)
URL: http://dergipark.gov.tr/humder

Alcohol User Prediction With Deep Learning Methods From Electronic Health Record Data

Elektronik Sağlık Kaydı Verilerinden Derin Öğrenme Yöntemleri ile Alkol Kullanıcısı Tahmini

Yazar(lar) (Author(s)): *Yasin KARAKUŞ*[1]

[1] ORCID ID: 0000-0002-4534-0151

Bu makaleye şu şekilde atıfta bulunabilirsiniz (To cite to this article): Karakuş Y., "Alcohol User Prediction With Deep Learning Methods From Electronic Health Record Data", ***Harran University Journal of Engineering***, 10(2): 94-104, (2025).

DOI: 10.46578/humder.1662856

Harran Üniversitesi Mühendislik Dergisi

Harran University Journal of Engineering

https://dergipark.org.tr/tr/pub/humder

Araştırma Makalesi (Research Article)

Alcohol User Prediction with Deep Learning Methods from Electronic Health Record Data

Yasin KARAKUŞ[1]*

[1]Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Kütahya Health Sciences University, Kütahya, Türkiye.

Article Information

Received: 21/03/2025
Revised: 22/04/2025
Accepted: 09/05/2025
Published: 30/06/2025

Citation

Karakuş Y., "Alcohol User Prediction With Deep Learning Methods From Electronic Health Record Data", Harran University Journal of Engineering, 10(2): 94-104, (2025).

Abstract

Alcohol consumption has negative effects on individuals and societies in various areas, including health, economic, social and cultural aspects. Alcohol use prediction is a very important research topic to prevent the negative effects of alcohol. While dose-dependent alcohol use disorder is usually predicted in the literature, in this study, unlike the literature, dose-independent alcohol users are predicted. This prediction is made from electronic health record data using popular deep learning methods. The dataset used in the study consists of 24 different attributes including personal characteristics and health parameters of 991346 individuals collected from the National Health Insurance Service in Korea. The data were optimised after digitisation and normalisation preprocessing steps. A certain amount of training and test separation was applied to the dataset. Then, an alcohol user prediction model was developed using artificial neural networks, LSTM and CNN method. According to the results obtained, although the models achieved close prediction success, artificial neural networks achieved the best result. After artificial neural networks, CNN ranked second, and LSTM ranked last. By using more than one deep learning method together in the study, a conclusion about the general success of deep learning methods on the current problem has been made and a method that will make an important contribution to the solution of the problem has been put forward.

Keywords: *Alcohol user, Deep learning, LSTM, CNN, ANN.*

Elektronik Sağlık Kaydı Verilerinden Derin Öğrenme Yöntemleri ile Alkol Kullanıcısı Tahmini

Makale Bilgisi

Başvuru: 21/03/2025
Düzeltme: 22/04/2025
Kabul: 09/05/2025
Yayınlanma: 30/06/2025

Alıntı

Karakuş Y., "Elektronik Sağlık Kaydı Verilerinden Derin Öğrenme Yöntemleri ile Alkol Kullanıcısı Tahmini", Harran Üniversitesi Mühendislik Dergisi, 10(2): 94-104, (2025).

Öz

Alkol tüketiminin bireyler ve toplumlar üzerinde sağlıksal, ekonomik, sosyal ve kültürel yönler de dâhil olmak üzere çeşitli alanlarda olumsuz etkileri vardır. Alkol kullanımının öngörülmesi, alkolün olumsuz etkilerini önlemek için çok önemli bir araştırma konusudur. Literatürde genellikle doza bağlı alkol kullanım bozukluğu tahmin edilirken, bu çalışmada literatürden farklı olarak dozdan bağımsız alkol kullanıcısı tahmini yapılmaktadır. Bu tahmin popüler derin öğrenme yöntemleri kullanılarak elektronik sağlık kaydı verilerinden yapılmaktadır. Çalışmada kullanılan veri kümesi, Kore'deki Ulusal Sağlık Sigortası Hizmetinden toplanan 991346 bireye ait kişisel özellikler ve sağlık parametrelerini içeren 24 farklı öznitelikten oluşmaktadır. Veriler, sayısallaştırma ve normalizasyon ön işleme adımlarından sonra optimize edilmiştir. Veri kümesine belirli miktarda eğitim ve test ayrımı uygulanmıştır. Ardından, yapay sinir ağları, LSTM ve CNN yöntemi kullanılarak bir alkol kullanıcısı tahmin modeli geliştirilmiştir. Elde edilen sonuçlara göre modeller birbirine yakın tahmin başarısı elde etse de en iyi sonucu yapay sinir ağları elde etti. Yapay sinir ağlarından sonra CNN ikinci sırada, LSTM ise son sırada yer aldı. Çalışmada birden fazla derin öğrenme yöntemi bir arada kullanılarak derin öğrenme yöntemlerinin mevcut problem üzerindeki genel başarısı hakkında bir sonuca varılmış ve problemin çözümüne önemli katkı sağlayacak bir yöntem ortaya konulmuştur.

Anahtar Kelimeler: *Alkol kullanıcısı, Derin öğrenme, LSTM, CNN, ANN.*

*İletişim yazarı, e-mail: yasin.karakus@ksbu.edu.tr
*Corresponding author, e-mail: yasin.karakus@ksbu.edu.tr

DOI: 10.46578/humder.1662856

1. INTRODUCTION

Alcohol, a psychoactive substance, can induce a state of intoxication that can have detrimental consequences on the human organism. The extent of these deleterious effects is contingent on the quantity and type of alcohol consumed, with higher doses resulting in more significant harm. In the short term, alcohol use can precipitate falls and injuries, as well as an elevated risk of violence, alcohol poisoning, and miscarriage or stillbirth in women. In the long term, it can inflict considerable harm on individuals and societies in areas such as health, psychology, and social life, with consequences including elevated blood pressure, heart diseases, paralysis, liver disorders, digestive problems, skin diseases, eye diseases, weakened immunity, an increased rate of cancer and difficulties in work, family and educational life. Research has demonstrated that alcohol consumption has been responsible for more than 140,000 deaths and 3.6 million years of potential life lost in the United States between 2015 and 2019, resulting in an average lifespan reduction of 26 years for those who have died [1]. Moreover, the economic cost of alcohol consumption in 2010 was estimated to be 249 billion dollars [2]. Preventative measures targeting alcohol consumption have been identified as a potential solution to mitigate the associated harms, particularly in terms of public health and the national economy. The process of detecting alcohol use is recognized as a crucial aspect in the formulation of preventive policies and the enhancement of awareness initiatives. While various methods exist for detecting alcohol use, data science methods have emerged as a particularly effective approach, offering the potential to reduce time and labor costs. Among the various data science methods, deep learning methods have garnered significant attention. The increasing prevalence of data has led to a notable rise in the utilization of deep learning methods, owing to their superior performance in addressing problems involving large data sets and generating more effective outcomes than conventional machine learning approaches. In essence, deep learning methods represent a class of machine learning tools that emulate the fundamental principles of human brain function.

1.1 Related Works

Machine learning methods and deep learning methods, which are sub-instruments of machine learning, have recently started to be used frequently in problems such as predicting diseases from symptoms by enabling a better understanding of development and treatment in biomedical research and have started to produce successful results. Mumtaz et al. [3] carried out a study to automatically identify alcohol use disorder from EEG. The EEG data utilized in this study was collected from 12 alcohol abusers during 10 minutes of eyes closed and eyes open conditions, and the principal component analysis (PCA) was employed to select the most relevant quantitative electroencephalography features. The performance of various machine learning models, including Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Multilayer Back Propagation Network (MLP), and Logistic Model Trees (LMT), was evaluated using 10-fold cross-validation (10-CV). Among these models, LMT demonstrated the most optimal outcomes, achieving classification accuracy 96%, sensitivity 97%, and specificity 93% metrics. The study also revealed significant neurophysiological distinctions between individuals with alcohol dependence, alcoholics, and the control group.

Ebrahimi et al. [4] proposed a method to predict AUD from electronic health record (EHR) data. The EHR data used in this study consisted of data from 2,571 patients aged between 18 and 101 years in the Southern Denmark Region. Firstly, they applied preprocessing to the EHR data, and then a feature selection method was applied to select the most relevant EHR features. Following these steps, the two main feature sets were explained, and the SVM classifier was tested on these two feature sets. The results showed that the SVM classifier achieved 0.80 and 0.92 accuracy, respectively.

Sisodia et al. [5] conducted a study to prevent the negative effects of alcohol on academic performance, evaluating classification algorithms to predict some of the risks of alcohol consumption among middle school students. The attributes in the dataset used in the study consist of grades of middle school students, and social, demographic, and school-related variables. Three distinct classifiers, namely Naive Bayes Classifier, Random Tree, and Simple Logistic, in conjunction with three ensemble classifiers, encompassing Random Forest, Bagging, and Adaboost, were evaluated on the dataset. The Random Forest approach yielded the most optimal outcome, attaining a ROC score of 0.981.

Kinreich et al. [6] conducted a study to classify individuals before the onset of alcohol use disorder (AUD) into two groups: those who developed AUD years later and those who did not. The study incorporated electroencephalography (EEG) measurements, family history information, and data on alcohol consumption, alcohol dependence, and a set of Genome-wide Association Study (GWAS) single-nucleotide polymorphisms (SNPs) from recent Genome-wide Association Studies (GWAS) of alcohol-related EEG measurements as features. An SVM model was then employed to analyze the feature set. The results of the study highlighted the importance of uniform sampling, followed by stratified analysis (e.g. based on ancestry, sex, and developmental stage) and broader feature selection to enable more accurate prediction of AUD development.

Dhillon et al. [7] proposed an Internet of Things (IoT)-based enterprise health information system, termed IoTPulse, for predicting alcohol dependence. This system provides real-time data using machine learning in a fog computing environment and was tested using data from 300 alcohol addicts from Punjab, India. Bayesian networks, neural networks, and k-nearest neighbor (kNN) models were tested on this data set. The Neural Networks model achieved an accuracy of 96.72% on the dataset collected for alcohol dependence prediction for IoTPulse.

In addition to studies in which the target variable is alcohol use, there are also studies in which alcohol use is used to estimate the value of the target variable.

Narkbunnum and Wisaeng [8] conducted a study to predict whether students were depressed or not. The study utilized a combination of socio-demographic characteristics, internet addiction, and stress levels data, along with alcohol use disorder information, to develop a predictive model. Saxena et al. [9] sought to predict liver impairment through the utilization of blood test measurements believed to be sensitive to liver disorders potentially resulting from high-capacity alcohol consumption. Tseng et al. [10] conducted a study using alcohol use and demographic, behavioral, and salivary autoantibody levels data to predict high-risk cases of oral cavity squamous cell carcinoma. Patel et al. [11] developed a periodontal disease prediction model using large sample sizes, machine learning (ML) that can provide up-to-date information, and electronic dental record (EDR) data including alcohol use. Qiu et al. [12] sought to predict the risk of osteoporosis in 2024, utilizing a comprehensive dataset encompassing bone mineral density, demographic and clinical information, including alcohol consumption. Their feature importance analysis revealed alcohol use to be among the top 10 features contributing to the prediction of osteoporosis. Ahmad et al. [13] analyzed the effects of age, gender, sleep duration, Rapid Eye Movement (REM) sleep percentages, deep sleep, light sleep, awakenings, caffeine consumption, alcohol consumption, smoking status, and exercise frequency on sleep efficiency. Their analyses indicated that higher alcohol consumption and smoking status were associated with lower efficiency. Unlu and Subasi [14] developed models with the objective of predicting the use of cannabis, ecstasy, amphetamine, cocaine, and non-prescription drugs in 2024. The models were developed using various data from the 2022 Finnish National Drug Survey, including alcohol use data from 3,857 respondents. The models used long short-term memory (LSTM), BiLSTM, and Recursive LSTM deep learning models to generate predictions. The findings yielded insights into the factors influencing substance use.

1.2 Observations and Contributions

Existing studies on alcohol estimation treat alcohol use disorder as a standalone entity, ignoring the fact that alcohol, being an addictive substance, can cause more use with each use, regardless of the dose. In order to prevent these effects, and to be used in the policies to be developed by individuals and institutions, in this study, people are classified as users or non-users, regardless of the dose of use, taking into account the above-mentioned effects. The study employs artificial neural networks (ANNs), LSTMs, and convolutional neural networks (CNNs) as deep learning methods for classification. The success of these algorithms in predicting alcohol users is then compared. The motivation behind the selection of these deep learning methods is that they are among the most preferred methods in deep learning problems with relatively different data structures. By using and comparing these methods in the prediction of alcohol users, it is aimed to provide a guide for the selection of deep learning methods that have different advantages over each other in terms of method selection for the current problem.

Initially, a dataset was meticulously prepared by the National Health Insurance Service in Korea, where personal information and sensitive data were extracted, to ensure the efficacy of the models in predicting alcohol users. The success of the trained models in predicting alcohol users is then evaluated using evaluation metrics. The study is divided into four sections. Firstly, an introductory section provides basic information about alcohol and alcohol use, defines the problem, and includes a literature review. This is followed by the materials and methods section, where the dataset, data preprocessing steps, and the deep learning-based alcohol user prediction model including ANN, CNN, and LSTM are explained. The subsequent section presents the results of testing the performance of the trained models with performance metrics. Finally, the study is concluded with a discussion of the study's findings and recommendations for future research.

2. MATERIAL and METHOD

2.1 Dataset

The dataset utilized in this study was obtained from the National Health Insurance Service in Korea, where sensitive data is extracted along with personal information. The dataset was obtained from Kaggle [15]. The dataset under scrutiny comprises 24 distinct columns and 991,346 rows, featuring body signals from smokers and alcohol users. The detailed composition of the dataset is outlined in Table 1. The correlation heat map of the dataset is illustrated in Figure 1. Distribution of the Target Variable by Gender is illustrated in Figure 2.

Table 1. Detailed Information About Dataset.

Feature Name	Description
Sex	Male or female
Age	Round up to 5 years
Height	Round up to 5 cm
Weight	Kg
Waistline	Cm
Sight_left	Eyesight (left)
Sight_right	Eyesight (right)
Hear_left	Hearing left, 1 (normal), 2 (abnormal)
Hear_right	Hearing right, 1 (normal), 2 (abnormal)
SBP	Systolic blood pressure (mmHg)
DBP	Diastolic blood pressure (mmHg)
BLDS	BLDS or FSG (fasting blood glucose) (mg/dL)
Tot_chole	Total cholesterol (mg/dL)
HDL_chole	HDL cholesterol (mg/dL)
LDL_chole	LDL cholesterol (mg/dL)
Triglyceride	Triglyceride (mg/dL)
Hemoglobin	Hemoglobin (g/dL)
Urine protein	Protein in urine, 1 (-), 2 (+/-), 3 (+1), 4 (+2), 5 (+3), 6 (+4)
Serum creatinine	Serum (blood) creatinine (mg/dL)
SGOT_AST	SGOT (Glutamate-oxaloacetate transaminase) AST (Aspartate transaminase) (IU/L)
SGOT_ALT	ALT (Alanine transaminase) (IU/L)
Gamma GTP	Y-glutamyl transpeptidase (IU/L)
SMK_stat_type_cd	Smoking state, 1 (never), 2 (used to smoke but quit), 3 (still smoke)
DRK_YN	Drinker or not
Sex	Male or female

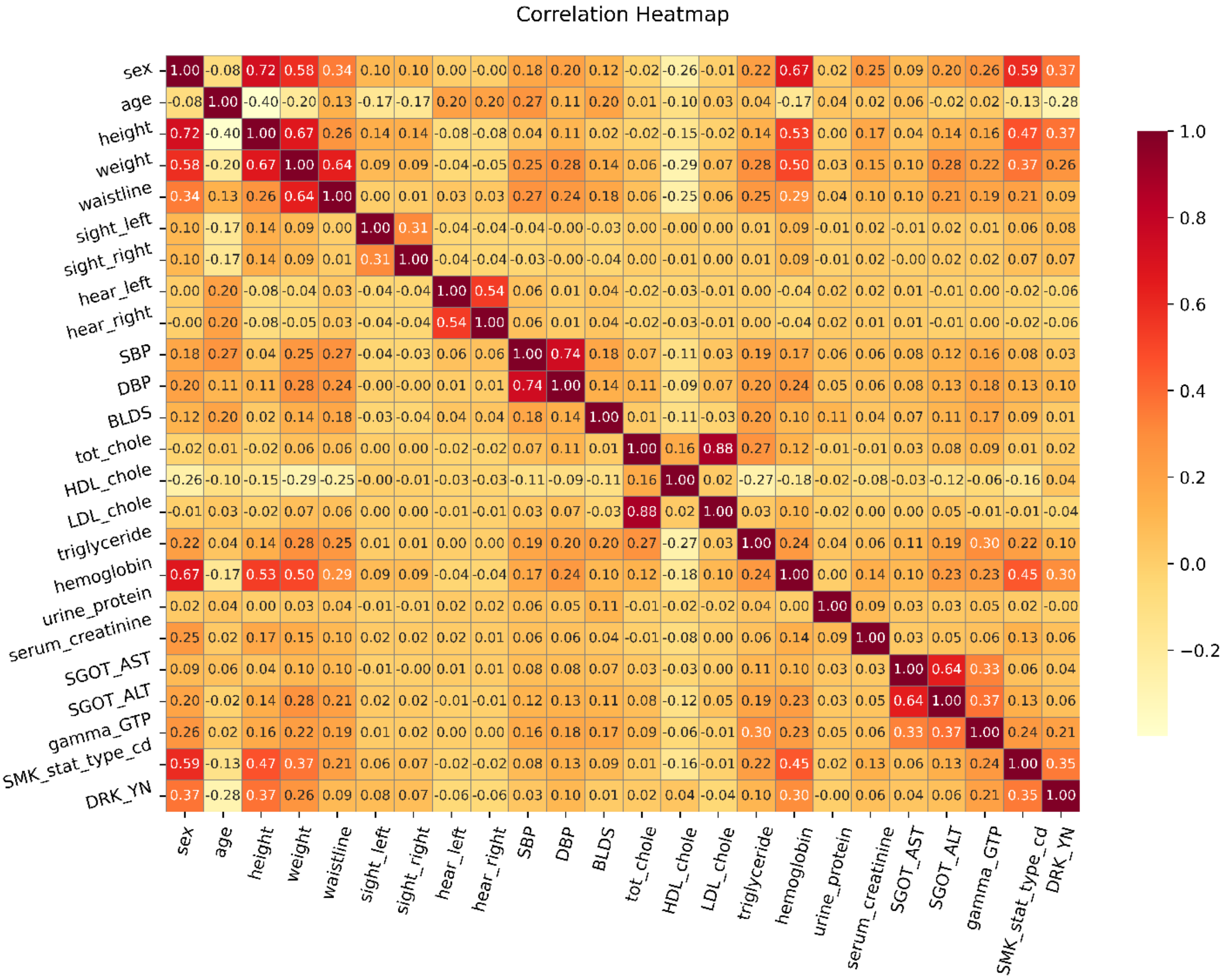


Figure 1. Correlation Heatmap of The Dataset.

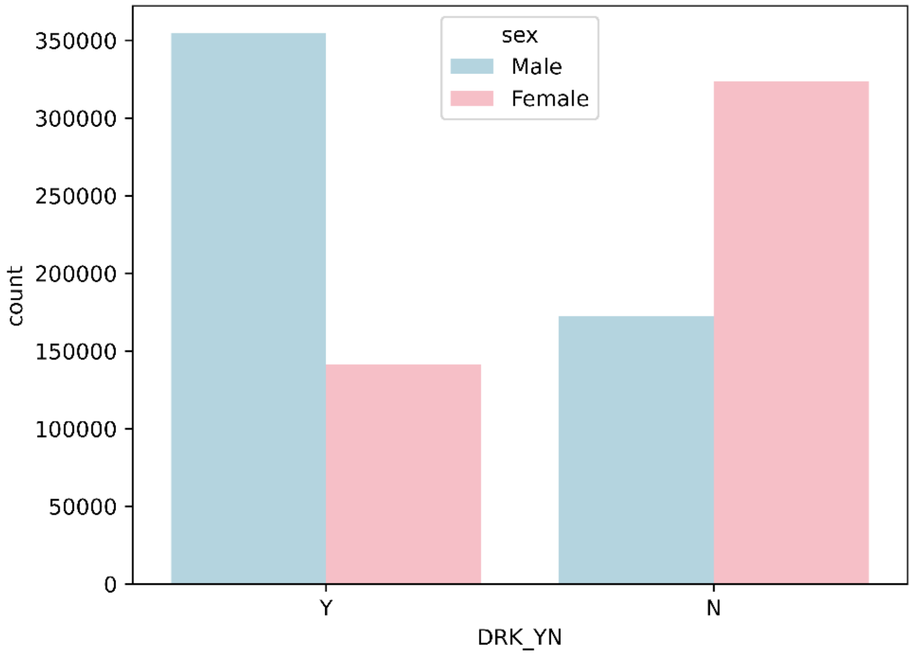


Figure 2. Distribution Of Target Variable by Gender.

2.2 Data Preprocessing

The data preprocessing stage is the stage where the data is processed and optimized to produce results. In this study, the data preprocessing stage consists of two different steps: label encoding and normalization. For learning methods to be effective on data, the data columns must be numeric. Label encoding is the process of representing textual data with simple numerical expressions. In our dataset, label encoding was applied to the 'sex' and 'DRK_YN' features.

Normalization is the process of placing data into smaller intervals to reduce computational costs and use memory and time-space more efficiently. The min-max normalization method, which is one of the most frequently used data normalization methods, is a method in which all data in the column are assigned to values between 0 and 1 after accepting the minimum value of each column in the data set as 0 and the maximum value as 1.

2.3 Alcohol User Prediction Model

After the label encoding and data normalization steps, the dataset was divided into two parts, 70% train and 30% test. The classifiers were then trained on the train dataset and finally tested on the test dataset and the performance of the classifiers was calculated with performance metrics. Models were tested independently of other models under the same conditions. The result of one model does not affect the other model. Parameters not specifically mentioned in the models are used with their default values.

The work was implemented in the JupyterLab 4.2.5 IDE using python 3.12.7 Successfully executed on a computer with Intel Core i9-13950HX 2.20 GHz CPU and 64 GB RAM.

2.3.1 Artificial neural network

Artificial neural networks are a computing technology adapted from the neurons in the human brain and the communication of neurons with each other. These networks represent a sophisticated approach to understanding the relationship between data input and output and are a form of deep learning that has been employed in the solution of numerous problems. The architecture of artificial neural networks comprises three distinct layers. These layers are the input layer, the hidden layer, and the output layer [6]. While the input and output layers comprise a single layer, the hidden layer can be subdivided into multiple layers. Each layer is capable of containing multiple neurons. The input and output layers contain vectors that comprise data about the input data and the inferences obtained from it. The function of the hidden layer is to facilitate the processes and operations required to derive the output from the input data. The feed-forward artificial neural network model developed for alcohol user prediction comprises three distinct hidden layers. The number of neurons in each of these layers is 64, 32, and 16, respectively. The artificial neural network model used in the study is illustrated in Figure 3. In ANN, Adam was selected as the optimizer, binary crossentropy as the loss function, relu as the activation function for the first, second and third ANN layer and sigmoid as the activation function for the dense layer.

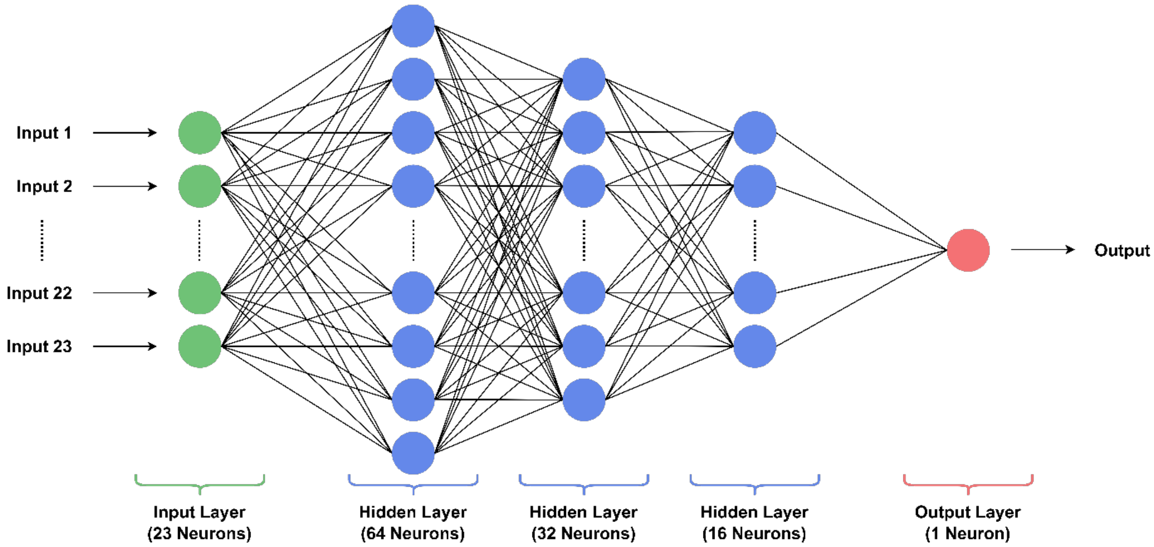
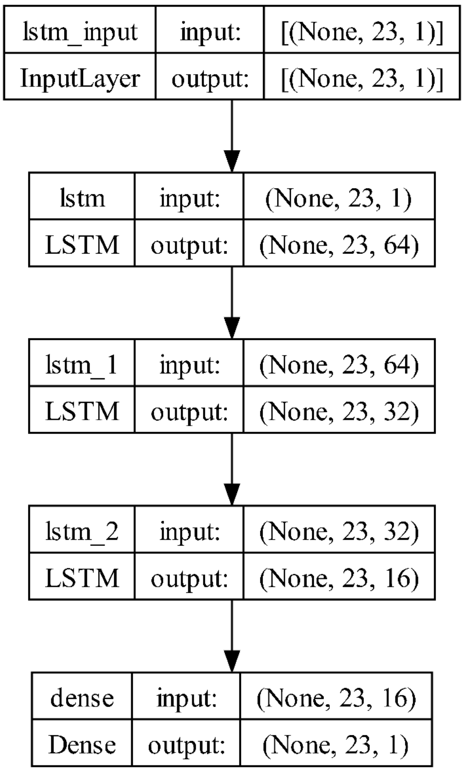


Figure 3. The Structure of ANN Used in The Study.

2.3.2 Long short-term memory

LSTM is an advanced recurrent neural network architecture that has gained popularity in the field of deep learning. It possesses the capacity to process not only instantaneous data, such as images but also sequences of data, including speech and video. A distinguishing feature of LSTM is its incorporation of feedback connections, which distinguishes it from standard feed-forward neural networks. This architectural distinction enables LSTM to circumvent the vanishing and exploding gradient problems that plague traditional recurrent neural networks. Consequently, LSTM exhibits superior proficiency in the inference of sequential, interrelated data. The LSTM model used in the study is illustrated in Figure 4. In LSTM, Adam was selected as the optimizer, binary crossentropy as the loss function, relu as the activation function for the first, second and third LSTM layer and sigmoid as the activation function for the dense layer.



lstm_input	input:	[(None, 23, 1)]
InputLayer	output:	[(None, 23, 1)]

lstm	input:	(None, 23, 1)
LSTM	output:	(None, 23, 64)

lstm_1	input:	(None, 23, 64)
LSTM	output:	(None, 23, 32)

lstm_2	input:	(None, 23, 32)
LSTM	output:	(None, 23, 16)

dense	input:	(None, 23, 16)
Dense	output:	(None, 23, 1)

Figure 4. The Structure of LSTM Used in The Study.

2.3.3 Convolutional neural network

Although convolutional neural networks are mostly preferred in image and video processing applications due to their success in recognizing patterns in images, they are also a preferred neural network architecture in other applications. Convolutional layers consist of 5 different basic components: activation functions, pooling layers, fully connected layers, and dropout layers [16]. The CNN model used in the study is illustrated in Figure 5. In CNN, Adam was selected as the optimizer and binary crossentropy was selected as the loss function. For the first, second and third CNN layer, the activation function relu, kernel size 3 was selected. In the pooling layer, pool size was selected as 3.

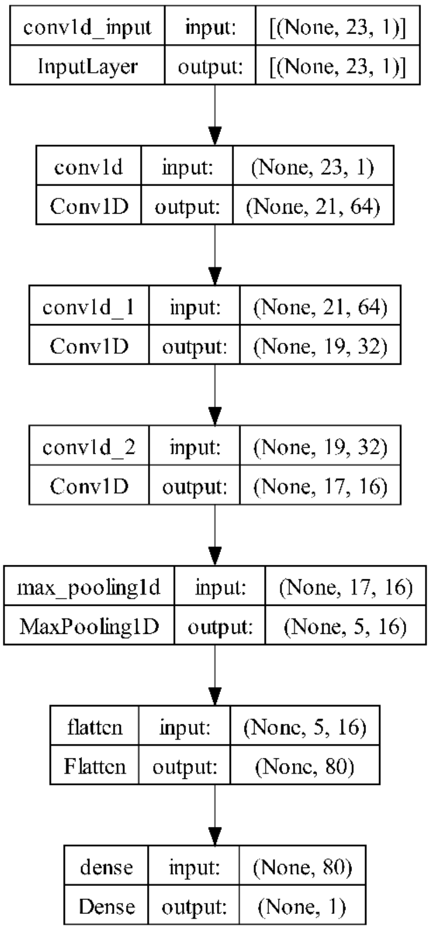

conv1d_input | input: | [(None, 23, 1)]
InputLayer | output: | [(None, 23, 1)]
conv1d | input: | (None, 23, 1)
Conv1D | output: | (None, 21, 64)
conv1d_1 | input: | (None, 21, 64)
Conv1D | output: | (None, 19, 32)
conv1d_2 | input: | (None, 19, 32)
Conv1D | output: | (None, 17, 16)
max_pooling1d | input: | (None, 17, 16)
MaxPooling1D | output: | (None, 5, 16)
flatten | input: | (None, 5, 16)
Flatten | output: | (None, 80)
dense | input: | (None, 80)
Dense | output: | (None, 1)

Figure 5. The Structure of CNN Used in The Study.

3. RESULTS

In this study, the accuracy, precision, recall, and F1-score metrics were utilized to evaluate the trained models. The formulae below are to be employed with the following definitions: TP is the number of correct predictions of the person who consumes alcohol. TN is the number of correct predictions of the person who does not consume alcohol. FP is the total number of predictions in which the person who consumes alcohol is predicted as a person who does not consume alcohol. Finally, FN denotes the number of predictions in which the person who does not drink alcohol is predicted as the person who drinks alcohol. Accuracy (ACC) is calculated according to Equation 1. Precision (PR) is calculated according to Equation 2. Recall (RE) is calculated according to Equation 3. F1-score (F1) is calculated according to Equation 4.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$PR = \frac{TP}{TP + FP} \tag{2}$$

$$RE = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 * \frac{PR * RE}{PR + RE} \tag{4}$$

In the evaluation results, it is seen that although the models have achieved close success scores for each other, the best score is given by artificial neural networks with a prediction accuracy of over 70% for each metric used. While CNN came after ANNs in the best prediction success, LSTM came last with a prediction accuracy below 70%. The evaluation results are shown in Table 2.

Table 2. Scores Obtained by Deep Learning Models in Test Results According to Evaluation Criteria.

Methods	Accuracy	Precision	Recall	F1
ANN	**0.7377**	**0.7380**	**0.7377**	**0.7376**
LSTM	0.6844	0.6851	0.6844	0.6841
CNN	0.7124	0.7125	0.7124	0.7123

4. CONCLUSION

Alcohol consumption is recognized to have detrimental consequences for individuals and communities, impacting their health, social relations, and cultural dynamics. In response to this recognized problem, numerous studies have been conducted to mitigate the adverse effects of alcohol. In these studies, dose-related alcohol use disorder is generally predicted. In this study, unlike other studies, dose-independent alcohol users were predicted. A significant body of these studies employs data science methodologies. This study aims to contribute to the broader effort to curtail alcohol consumption by developing a prediction model using EHR data. The EHR data contains biometric signals, from which personal and sensitive information is extracted through the utilization of deep learning methods, namely ANN, LSTM, and CNN. In this study, three distinct deep learning methods were utilized to assess the efficacy of deep learning algorithms in addressing the prevailing issue. The findings revealed that while all models demonstrated notable predictive accuracy, ANNs emerged as the most effective approach. After ANN, CNNs were positioned as the second most successful method, while LSTMs ranked least effective. The integration of multiple deep learning methods in the present study has enabled the formulation of a comprehensive conclusion regarding the efficacy of these methods in addressing the prevailing issue. Furthermore, a methodology that is poised to make a substantial contribution to the resolution of the problem has been proposed.

In subsequent studies, the scope will be expanded to encompass datasets with a greater number of features and a more diverse range of data categories, thereby facilitating the investigation of a more extensive array of areas where alcohol is either the cause or the consequence. The efficacy of alternative deep learning methods can be assessed through self-assessment or by comparing them with traditional machine learning methods.

CONFLICT OF INTEREST

The author/authors declare that there are no conflicts of interest regarding the publication of this article.

STATEMENT OF PUBLICATION ETHICS

I declare that all processes of the study comply with research and publication ethics, and that I comply with ethical rules and scientific citation principles.

AUTHOR STATEMENT

Yasin Karakuş: Idea/Concept, Design and Conception, Supervision, Consultancy, Resources/Materials, Data Collection, Data Processing, Analysis/Interpretation, Literature Review, Drafting, Layout, Writing the Main Text.

REFERENCES

[1] Centers for Disease Control and Prevention. (n.d.). *Alcohol screening and brief intervention (SBI)*. Centers for Disease Control and Prevention. https://www.cdc.gov/alcohol-pregnancy/hcp/alcoholsbi/index.html

[2] Sacks, J. J., Gonzales, K. R., Bouchery, E. E., Tomedi, L. E., & Brewer, R. D. (2015). 2010 national and state costs of excessive alcohol consumption. *American Journal of Preventive Medicine*, *49*(5). https://doi.org/10.1016/j.amepre.2015.05.031

[3] Mumtaz, W., Vuong, P. L., Xia, L., Malik, A. S., & Rashid, R. B. (2016). Automatic diagnosis of alcohol use disorder using EEG features. *Knowledge-Based Systems*, *105*, 48–59. https://doi.org/10.1016/j.knosys.2016.04.026

[4] Ebrahimi, A., Wiil, U. K., Andersen, K., Mansourvar, M., & Nielsen, A. S. (2020). A predictive machine learning model to determine alcohol use disorder. *2020 IEEE Symposium on Computers and Communications (ISCC)*, 1–7. https://doi.org/10.1109/iscc50000.2020.9219685

[5] Sisodia, D. S., Agrawal, R., & Sisodia, D. (2018). A comparative performance of classification algorithms in predicting alcohol consumption among secondary school students. *Advances in Intelligent Systems and Computing*, 523–532. https://doi.org/10.1007/978-981-13-0923-6_45

[6] Kinreich, S., Meyers, J. L., Maron-Katz, A., Kamarajan, C., Pandey, A. K., Chorlian, D. B., Zhang, J., Pandey, G., Subbie-Saenz de Viteri, S., Pitti, D., Anokhin, A. P., Bauer, L., Hesselbrock, V., Schuckit, M. A., Edenberg, H. J., & Porjesz, B. (2019). Predicting risk for alcohol use disorder using longitudinal data with multimodal biomarkers and family history: A machine learning study. *Molecular Psychiatry*, *26*(4), 1133–1141. https://doi.org/10.1038/s41380-019-0534-x

[7] Dhillon, A., Singh, A., Vohra, H., Ellis, C., Varghese, B., & Gill, S. S. (2020). IoTPulse: Machine learning-based Enterprise Health Information System to predict alcohol addiction in Punjab (India) using IOT and fog computing. *Enterprise Information Systems*, *16*(7). https://doi.org/10.1080/17517575.2020.1820583

[8] Narkbunnum, W., & Wisaeng, K. (2022). Prediction of depression for undergraduate students based on imbalanced data by using data mining techniques. *Applied System Innovation*, *5*(6), 120. https://doi.org/10.3390/asi5060120

[9] Saxena, S., Deep, V., & Sharma, P. (2018). Liver disorder prediction due to excessive alcohol consumption using slave. *Advances in Intelligent Systems and Computing*, 193–202. https://doi.org/10.1007/978-981-13-1951-8_18

[10] Tseng, Y.-J., Wang, Y.-C., Hsueh, P.-C., & Wu, C.-C. (2022). Development and validation of machine learning-based risk prediction models of oral squamous cell carcinoma using salivary autoantibody biomarkers. *BMC Oral Health*, *22*(1). https://doi.org/10.1186/s12903-022-02607-2

[11] Patel, J. S., Su, C., Tellez, M., Albandar, J. M., Rao, R., Iyer, V., Shi, E., & Wu, H. (2022). Developing and testing a prediction model for periodontal disease using machine learning and Big Electronic Dental Record Data. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.979525

[12] Qiu, C., Su, K., Luo, Z., Tian, Q., Zhao, L., Wu, L., Deng, H., & Shen, H. (2024). Developing and comparing deep learning and machine learning algorithms for osteoporosis risk prediction. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1355287

[13] Ahmad, H., Umar Khan, M., & Azam, M. (2024). Comparative analysis of machine learning methods for enhancing sleep efficiency and prediction. *Information Systems Engineering and Management*, 3–15. https://doi.org/10.1007/978-3-031-66854-8_1

[14] Unlu, A., & Subasi, A. (2024). Substance use prediction using artificial intelligence techniques. *Journal of Computational Social Science*, *8*(1). https://doi.org/10.1007/s42001-024-00356-6

[15] Soo.Y. (2023, August 30). *Smoking and drinking dataset with body signal*. Kaggle. https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset

[16] Subramaniam, K., & Naganathan, A. (2024). Enhancing retinal fundus image classification through active gradient deep convolutional neural network and red spider optimization. *Neural Computing and Applications*, *36*(26), 16607–16619. https://doi.org/10.1007/s00521-024-09989-0