



## Detection and Prevention of Medical Fraud using Machine Learning

Ceyda ÜNAL<sup>1</sup> , Gökçe Sinem ERBUĞA<sup>2</sup> 

<sup>1</sup>Dokuz Eylül University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, İzmir, Türkiye

<sup>2</sup>Dokuz Eylül University, Faculty of Economics and Administrative Sciences, Department of Business Administration, İzmir, Türkiye

**Corresponding author** : Gökçe Sinem ERBUĞA

**E-mail** : gokce.erbuga@deu.edu.tr,  
gokceerbuga@gmail.com

\* This study is a revised and expanded version of the paper presented at the 22nd International Business Congress organized by Nişantaşı University on 07-09 September 2023.

### ABSTRACT

Presently, there is an upward trend in the mean life expectancy of individuals due to reductions in maternal and infant mortality, as well as deaths caused by non-communicable diseases like cardiovascular disease. A decline in life expectancy results in a corresponding increase in health expenditures sustained by both public and private entities, including insurance providers. The healthcare sector has become an extremely comprehensive and critical industry due to the following factors: the increase in healthcare expenditures, particularly during the pandemic; the cost of each component in the healthcare sector; the increasingly chaotic healthcare technology ecosystem; the growing expectations of numerous and diverse stakeholders; and the presence of numerous and new actors in the sector. Nevertheless, this circumstance exposes the health sector to many hazards, thereby increasing its susceptibility to fraudulent activities. The sector's substantial volume will inevitably lead to expensive fraudulent activities. For this reason, prospective medical frauds should be prevented and detected immediately. Machine learning is considered one of the most powerful and optimal approaches to prevent medical fraud. An example application is used to assess the efficacy of machine learning in the medical fraud detection context as part of the research. The objective of the proposed application is to classify provider-side medical fraud by applying various machine learning techniques and medical claims.

**Keywords:** Machine Learning, Artificial Intelligence, Healthcare Sector, Medical Fraud.

**Submitted** : 02.04.2024

**Revision Requested** : 17.07.2024

**Last Revision Received** : 21.08.2024

**Accepted** : 08.09.2024

**Published Online** : 16.09.2024



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

## 1. INTRODUCTION

Noncommunicable diseases (NCDs) cause the highest disease burden worldwide, and their effects have worsened over the last 20 years. NCDs, which are a significant threat to people from all geographical backgrounds and age groups, affect elderly individuals and genetically and psychologically disadvantaged individuals to a greater extent. In addition, such diseases can cause greater losses in countries where lower-income groups are the majority, as stated by the World Health Organization.

Non-communicable diseases (NCDs) have grown to be a contributory factor in more than 40 million deaths worldwide, representing an increase of more than 30% since 2000. In other words, approximately 75% of deaths worldwide are caused by non-communicable diseases (NCDs) (WHO, 2023a). Although these diseases are quite prevalent, the lengthy and expensive procedures required to treat them force nations to set aside enormous sums of money for medical expenses. Therefore, countries must allocate huge budgets for health expenditures.

Considering 2020 data, the country that allocated the highest budget for health expenditures in the world was the United States, which constituted approximately 16.8% of the US GDP. The United States is followed by England (12.8%), Germany (12.5%), France (12.4%), and Canada (11.6%). Considering OECD countries, the ratio of health expenditures to GDP is determined as 8.8%, while in Türkiye this rate is equal to 4.7% of GDP (Euronews, 2022).

When evaluated on a global basis, the importance of health expenditures for all countries increases because of high health expenditures and rigid supply elasticity in health services. For this reason, any act of corruption or fraud in the process of realizing health expenditures will cause the costs that both the state, private sector organizations, and individuals have to bear to increase many times.

The healthcare sector is further made more complex by the high level of uncertainty, the significant number of players involved, the presence of asymmetric knowledge among these actors, and numerous other considerations. This complex structure makes the sector more open to corruption and fraud and paves the way for more illegal tasks to be undertaken in the industry (Avcı & Teyyare, 2012:199). Making the best use of the resources already available is crucial for ensuring the greatest level of effectiveness from health facilities offered in the health sector, where expenses are fairly high and unavoidable (tight flexibility) (Aydın & Yenimahalleli Yaşar, 2020:64). In this way, more people will be able to access health care services, and these people will be able to receive them inexpensively and efficiently. As a result, it is crucial to identify and prevent corruption and fraud that will disrupt the balance of optimal resource use in healthcare.

In this regard, being able to detect and prevent fraud that has occurred or will occur in the healthcare sector is an indispensable part of healthcare management and medical accounting. However, the large number of participants in the health sector and the existence of a large amount of data make it challenging to analyze and interpret such data using classical data analytics methods. In addition, it is not possible to store the obtained data and preserve them or use them effectively for diagnostic purposes during the treatment process using traditional methods. For this reason, developments in information and communication technologies are frequently used to collect, store, analyze, and interpret existing large volumes of data (Kurşun, 2021:921; Altındış & Kıran Morkoç, 2018:257). Likewise, technology has begun to be frequently used to detect medical frauds occurring in the healthcare sector. One of the methods used to prevent medical fraud is to detect fraud using machine learning. With machine learning, medical fraud can be prevented more effectively and at lower cost, and data in healthcare services can be interpreted more effectively and efficiently.

Within the scope of the study conducted in this regard, the concept of medical fraud was first explained in detail. In the following section, the detection of medical fraud is examined, followed by research on the topic published in the literature. After discussing how and in what way machine learning is used to detect medical fraud, the practical application of machine learning to detecting medical fraud is examined in an application case. Finally, the inferences drawn from the application were discussed, and necessary suggestions were made in the Conclusion.

## 2. CONCEPTUAL FRAMEWORK

Spending on healthcare is rising significantly because of population longevity and intensified access to healthcare globally. Due to the nature of the health sector, resources allocated to this sector have a large share in global and domestic economies. For example, healthcare expenditures in the United States reached 3.65 trillion USD or 11,172 USD per capita in 2018 (Ekin et al., 2021:1183). The uncertainties and risks in this sector are excessive, and there are many actors and imbalances between their knowledge about the sector (Avcı & Teyyare, 2012:199). This situation makes supervision of the health sector quite difficult. Therefore, the sector offers a more suitable playground for corruption than many other sectors.

Medical fraud creates a huge burden on both the health systems and economies of countries. Although an exact

amount cannot be determined, expenses due to healthcare fraud, forgery, and abuse in the medical sector can reach one-tenth of all healthcare expenditures. This rate is approximately \$100 billion in US health expenditures, which exceed \$1 trillion each year (US Department of Justice, 2024). In addition to the financial burden it imposes, fraud increases the cost of healthcare services and causes a huge decrease in their quality.

Because of medical fraud cases, \$500 billion the \$7.5 trillion annually spent on health care worldwide is wasted because of systemic corruption. According to another source, the global health system has reduced 6.2% of its average annual revenue because of corruption and fraudulent activities (Bozhenko, 2022:32). Although losses incurred because of medical fraud activities averaged \$1,297,560, 1-5% of these activities resulted in losses of \$150,000 or less, according to the report of the US Sentencing Commission (USSC). The average loss caused by criminal activities was \$9,500,000 (USSC, 2022).

To evaluate the situation within European borders, the damage caused by fraud and corruption cases in the healthcare sector is more than €50 billion. As shown in the report titled "The Financial Cost of Healthcare Fraud" prepared by Gee, Button, and Brooks from the Center for Counter Fraud Studies at the University of Portsmouth, EU countries spend more than €1 trillion on the health sector each year. €56 billions of this expenditure is wasted each year as losses due to corruption. The waste generated in the health sector worldwide can reach up to €180 billion.

This amount corresponds to more than 5% of the budget allocated by these countries to health expenditures (Vincke & Cylus, 2011:14). On the other hand, the annual amount of health funds withheld from the use of the health system each year is equal to the amount required to achieve Universal Health Coverage (UHC), the core commitment of the UN's Sustainable Development Goals, which aim to provide affordable, accessible, and quality health care for all (Transparency International, 2023).

The National Health Care Anti-Fraud Association (NHCAA) stated that economic losses from medical fraud are in the tens of billions of USD each year. Another optimistic estimate by the NHCAA states that the cost of medical fraud is 3% of the total amount of money spent on medical care in the US. However, the cost of medical fraud can reach up to 10% of annual healthcare expenditures, according to some government and law enforcement agencies, representing approximately more than 300 billion USD (NHCAA, 2021). In addition, it is estimated by Transparency International (2023) that fraud within the field of medical care causes an estimated 140,000 pediatric fatalities each year.

When we look at China, one of the largest economies in the world, it is seen that it ranks 5th in the health and health systems ranking, and the economic value of the health sector exceeds 2 trillion yuan. In this sector, which is a huge market, China's National Health Care Security Administration (NHSA) randomly inspected approximately 200 000 medical institutions in 2018. Because of the inspections conducted, approximately 1/3 of these institutions committed medical fraud (Zhang et al., 2020:2). According to the 2020 data of the National Healthcare Security Administration (NHSA), approximately 10% of health expenditures in China were wasted with these abuses.

Medical fraud affects all healthcare professionals, institutions, local and national governments, development banks, aid organizations, etc. actively operating in the healthcare sector. It directly affects many actors. To achieve the goal of "ensuring healthy lives and promote well-being for all at all ages" (SDG-3), one of the Sustainable Development Goals set by the United Nations, it is imperative for all nations and health institutions to devise strategies that tackle corruption (Vian, 2020:114).

Healthcare organizations are exceptionally susceptible to medical fraud. This is due to the following factors: unpredictability regarding service demand (including who will become unwell and when, as well as what they will require), the complex interactions of numerous different scattered parties, such as payers, suppliers, customers, and regulators, asymmetric information between different actors, and difficulty in identifying and controlling different interests, etc. (Vian, 2007:84). It is extremely important to determine to what extent private providers should be entrusted with important public roles in the provision of health (medical) services. In addition, as Savedoff and Kussman (2006) declare, the large public budget allocated for medical expenses in many countries requires that medical expenses be made transparently and meticulously.

Medical fraud observed in healthcare companies finds its place in the literature as a particular form of white-collar crime used to express the dishonest provision of medical services for the purpose of economic gain, that is, profit (Ogunbanjo et al., 2014:10). White-collar crimes, which include acts of "deception, concealment, or abuse of trust" and are not related to actions involving physical force or violence/threats (FBI, 1989:3), are committed by individuals or organizations established for the purpose of committing crimes with the motivation of individuals or economic gain.

The European Healthcare Fraud & Corruption Network (EHFCN) defines fraudulent acts carried out in the healthcare sector as "obtaining any gain/benefit improperly by deliberately violating a certain rule". Corruption is explained as "the abuse of authority by involving a third party in a criminal act and, as a result, obtaining a certain gain illegally" (Küçük, 2022: 588).

Patients/taxpayers, suppliers (companies from which the government purchases drugs and equipment), regulators (ministry of health, pharmaceutical regulatory agency), and suppliers (medical facilities such as hospitals, health clinics, and individual or group physician practices) are the five key actors in health systems as identified by the Organization for Economic Co-operation and Development (OECD) (Couffinhall & Frankowski, 2017). Crimes resulting from breaches of integrity, such as fraud and corruption, manifest due to engagement among these diverse stakeholders (Vian, 2020:116).

The planning of medical fraud crime offenses can vary in complexity. These offenses can also be committed by patients, healthcare professionals, or other entities who purposefully mislead the healthcare system to obtain fraudulent benefits or compensation (Thomson Reuters, 2021). The people who commit the crime of medical fraud include doctors, nurses, pharmaceutical companies, pharmacists, health technicians, medical officers, physiotherapists, and other healthcare professionals. In addition, individuals who demand healthcare (patients, consumers), healthcare companies, insurance companies, or actors who play an intermediary role in service delivery (medical service and equipment suppliers) can also be perpetrators of medical fraud (Price and Norris, 2009: 286). These people perform different types of medical criminal acts in illegal and unethical ways.

Medical fraud is frequently performed in various ways, such as upcoding (charging for a more expensive diagnosis or procedure), providing needless treatments or screenings, paying for healthcare services that were never rendered, unbundling, different criminal acts include falsifying the seriousness of a medical disease, exaggerating, manipulating, and paying illegal kickbacks in exchange for preferential treatment (Vian, 2007). There are many types of fraud and corruption in the healthcare industry. Even though the literature contains an extensive amount of research that subjected these criminal acts to different classifications, it is not possible to discuss a common typological classification. In the report titled "Corruption in the Healthcare Sector" prepared by the European Commission (EC), acts of fraud that may occur in the healthcare sector are grouped under the following six headings:

- Bribery in providing medical services, and
- Medical equipment supply fraud,
- Improper marketing relationships,
- Abuse of (high-level) professional positions,
- Unnecessary refund requests,
- Abuses and corruption related to medicines and medical devices (EC, 2017:9).

Transparency International, on the other hand, classifies fraud in the health sector into five categories:

- Embezzlement, and theft from the health budget or user fees,
- Corrupt procurement practices,
- Corrupt practices in payment systems,
- Corrupt practices in the pharmaceutical supply chain and
- Corrupt practices in the provision and delivery of health services (Transparency International, 2006:18).

Based on the known identity of the offender, Küçük (2022) classified fraud and corruption offenses in the healthcare industry into three categories: actions committed by patients, actions carried out by healthcare providers, and actions committed by patient or supplier actors, in other words, by third parties. A similar classification was observed in Ekin et al. (2018). The study by Ekin et al. in question classifies the concept of medical fraud through the perpetrators of the action and discusses it in three basic categories: crimes committed by the service provider (hospital, doctor, etc.), crimes committed by the consumer (patient), and crimes committed by insurers.

It is extremely difficult to identify frauds that occur with the help of globalization in the health sector and developments in information technologies using traditional auditing methods. It is possible to say that new crime types are emerging, and current crime detection methods are insufficient to counter fraud. Therefore, to carry out an effective fraud detection process, more advanced and comprehensive crime prevention systems that incorporate different statistical approaches are required (Li et al., 2008:275).

In order to identify the abuses encountered in the healthcare sector, the existence of a beneficial audit and control system and its effective functioning must be ensured. In the process where the rapid transformation in information technologies has not yet clearly demonstrated their effects, audits and controls are performed by audit personnel, such as auditors, controllers, and inspectors, and require longer time and more effort. This classical audit and control process, in other words, a process performed manually, mostly involves the examination of physical documents and files and long-term field work (Turğay et al., 2020:5).

Traditional healthcare fraud detection methods, which are often limited to efforts to detect fraud rather than prevent

it, have not yet been sufficiently efficient and effective. Detecting receivables before payment has been shown by the International Social Security Association (ISSA) as a more effective way to prevent fraud and corruption in the healthcare sector. Therefore, the paradigm of inappropriate healthcare expenditure management is shifting from surveillance management to prevention (ISSA, 2022).

The widespread use of digitalization and the resulting social transformation in society have led to the need for audit activities on electronic platforms. In addition, the large volume of data used in audits and accumulated over the years makes it worthwhile for audit staff to examine these data. These difficulties encountered during the audit process reveal that audit activities should not be satisfied by classical methods and that data should be examined in a more reliable way using new analysis and modeling techniques. In this context, in line with current developments, the concept of continuous auditing, which refers to "a form of auditing that enables the auditing of real-time accounting information (data) without the need for physical documents and is carried out by using a number of special computer-based audit programs", has emerged (Orhan, 2015: 85).

Because medical fraud causes losses between 3% and 10% of countries' health expenditures (Shin et al., 2012:7441), it is extremely important to audit these transactions carefully. The intensity and complexity of criminal allegations that may lead to medical fraud and corruption necessitate frequent use of information technology and data analysis methods in audit processes (Ekin, 2019:4107; Turgay et al., 2020:6). In addition, data analysis and modeling techniques have gained importance in this field, resulting in the emergence of various data analysis techniques.

Although classical (traditional) methods, such as linear discriminant analysis or logistic regression analysis, which are used to reveal fraud and corruption, are effective in solving many cases, there are more powerful and effective analysis methods, such as artificial neural networks (Bolton and Hand, 2002:237). He, Wang, Graco, and Hawkins (1997) and He, Graco, and Yao (1999) used neural networks, genetic algorithms, and nearest neighbor methods in their studies. In their study, Major and Riedinger (1992) used statistical information-based medical fraud detection methods that compared observations with those that were most similar. Different methods, such as artificial intelligence, machine learning, distributed and parallel computing, econometrics, expert systems, fuzzy logic, outlier detection, pattern recognition, and visualization, have also been used in the literature (Bolton and Hand, 2002:245).

By means of data-driven inventiveness, fraud and corruption detection and prevention technological advances, including in the cognitive field of computing, data mining, analytics, machine learning, and various other kinds of artificial intelligence (AI), have greatly advanced. Some of these methods are (ISSA, 2022):

- Biometric recognition using a fingerprint scanner, iris recognition, or facial recognition
- Predictive modeling methods, such as data mining, predictive analytics, and quantitative analysis techniques, to detect patterns in supplier fraud and behavior
- Artificial intelligence-based pattern recognition techniques used to identify coding and billing errors
- Blockchain applications make it impossible for fraudulent practices to delete or modify data and allow detailed asset tracking.

The Centers for Medicare and Medicaid Services (CMS), an American health insurance provider, publishes the healthcare data used by most researchers to detect healthcare fraud. Raw data for detecting fraud in healthcare businesses often come from insurance claims, including government healthcare data, physician data, clinical data, and private insurance company data.

As the rate of digital transformation in the healthcare industry accelerates, healthcare organizations' digitalization processes have begun to take on a completely novel form. In healthcare organizations, the emergence of electronic health data of various sizes and types has created new opportunities for automated fraud detection. Specifically, in the context of automated processes, machine learning and data mining techniques are crucial for detecting abuse in such data (Joudaki et al., 2015). Currently, machine learning techniques are regarded as the most essential components for identifying inappropriate use and fraud. Data mining and machine learning, as a collective, encompass methodologies that leverage artificial intelligence, statistics, and mathematics to extract and discover valuable insights from databases (Aydoğan Duman & Sagioglu, 2017). According to Alpaydın (2020), machine learning refers to the capacity of computer algorithms to gain decision-making skill using data and statistical theory when constructing models. In essence, there are three classifications for machine learning methodologies: supervised, unsupervised, and semi-supervised learning.

Supervised learning methods are employed to identify fraudulent or dishonesty activities in healthcare organizations. These methods utilize data samples that have been previously identified as fraudulent or non-fraudulent or labeled as a result. Models developed using these data are crucial for automated identification of previously identified fraud and abuse patterns. Methods of supervised learning employ a variety of techniques, including classification and regression algorithms.

Support Vector Machines (Francis, Pepper, & Strong, 2011; Kirlidog & Asuk, 2012), Neural Networks (Liou, Tang, & Chen, 2008), and decision trees (Branting et al., 2016) are examples of supervised machine learning techniques used to classify fraud detection in healthcare organizations. In addition, regression analysis methods are also used in fraud detection (Francis, Pepper, & Strong, 2011). At the same time, some studies have utilized naive Bayes and decision trees for big data analysis on the Hadoop platform (Dora & Sekheran, 2015). In this context, the most critical drawback of supervised machine learning methods is the need for human input and the required output, i.e., labeled data. In particular, for fraud detection, the acquisition and interpretation of labeled data are laborious and time-consuming (Saravanan and Sujatha, 2018). Unsupervised learning techniques have been suggested to address these disadvantages of supervised learning.

Unsupervised learning approaches identify fraudulent activities within an unannotated dataset, operating under the general assumption that a substantial proportion of the data comprises legitimate activities (Abdallah, Maarof, & Zainal, 2016). In contrast to supervised learning, the proposed model is constructed without the use of labeled data. One of the primary benefits of unsupervised learning is that it enables the precise detection of fraudulent activity, even in the absence or presence of inadequate labeled data (Bolton and Hand, 2001). Basic unsupervised learning methods include clustering, association rules, and outlier detection. Different machine learning methods such as association rule analysis (Shan et al., 2008), k-means (Shan et al., 2009), probabilistic programming (Bauder & Khoshgoftaar, 2016), are used in fraud detection in healthcare.

The benefits of both supervised and unsupervised learning are combined in semi-supervised learning. It can be considered a hybrid method that uses the features of supervised and unsupervised learning to achieve more accurate results. When there is a relatively small proportion of labeled data with a large amount of unlabeled data, semi-supervised machine learning techniques are frequently used. Building models that consider both labeled and unlabeled input is, in essence, the primary objective of semi-supervised learning (Zu, Wang, & Wu, 2011). With the assistance of domain experts, semi-supervised learning is also used in unsupervised learning techniques such as clustering and outlier detection (van Capelleveen et al., 2016).

### 3. LITERATURE REVIEW

Acts of fraud and corruption in the healthcare sector deepen social inequality, and in this context, poor people and disadvantaged groups are affected the most. In addition, fraud and corruption prevent the fight against important diseases because of the diversion of resources and funds for financing health services through different channels. For this reason, fighting corruption and fraud in healthcare is extremely important across the world to supply outstanding medical care and effectively address both current and potential risks to global healthcare by making healthcare accessible to all (Transparency International, 2023).

The U.S. The Department of Justice reported in 2023 that a hospice medical director was sentenced to four years and two months in prison. This sentence was imposed because the director submitted over \$150 million in false and fraudulent Medicare claims for hospice and other medical services. According to court filings, the healthcare organization's initiatives enroll people with incurable diseases, including dementia and Alzheimer's disease, and those with low mental capacity in retirement, nursing, and public housing. Patients were informed that their prognosis was less than six months by the corporation. Over \$18 million in unnecessary services were approved (U.S. Department of Justice, 2023).

The Georgetown University Memory Disorders Program's recent paper claims that medical fraud has been moved to academic research. Recent Alzheimer's disease (AD) research fraud raises serious issues. Several figures in frequently cited 2006 Nature research on animal models of AD may have been modified, leading to problematic results (GU University Memory Disorders Program, 2023). People are prosecuted for Alzheimer's misdiagnosis.

Because of the problems it creates in the pharmaceutical supply chain, medical fraud also causes disruptions to individual treatment processes. Corruption also results in the waste of existing economic and human resources by limiting countries' capacity to manage national and global health risks, according to Transparency International (2023).

Fraud and corruption crimes committed in the healthcare sector, in addition to their huge financial losses, also threaten the quality and safe delivery of medical services that the healthcare system can offer to individuals (Li et al., 2008:275). In this regard, acts of medical fraud and corruption should be detected as soon as possible and even prevented before they are revealed to enhance the standards of the amenities provided and concurrently cut back on service expenses.

The World Health Organization (WHO) attach great importance to the anti-fraud and anti-corruption, transparency, and accountability (ACTA) steps to be taken in this direction. Within the framework of UN Sustainable Development

Goal (SDGs) 16.5, reducing and even preventing corruption and bribery is necessary to improve health services, to prevent inequalities in health services, and to improve lives (WHO, 2023b).

In conformity with the 2020 report by the Organization for Economic Co-operation and Development (OECD), over 45% of the global population holds the belief that the healthcare industry is highly corrupt. From this vantage point, one of the primary concerns regarding steady growth for both highly wealthy countries and developing nations globally is the fight against fraud and corruption in the health sector.

The literature highlights the US Health Care Financing Administration (HCFA) among prominent government health departments. A pair of healthcare programs known as "Medicare" and "Medicaid" exist in the United States. Medicare is a government-administered social insurance plan for those with end-stage renal illness, as well as those aged 65 years or above, or those with specific disabilities who are younger than 65 years. This program offers prescription drug coverage, hospital insurance, and health insurance. Medicaid is administered by individual states, with each state establishing its own eligibility and service requirements. Medicaid is restricted to low-income families and individuals who satisfy the eligibility requirements established by state and federal legislation (Liu & Vasarhelyi, 2013). Several data types are subjected to distinct analyses because health-system data vary by country.

Liou et al. (2008) examined claims submitted to the Taiwan National Health Insurance for outpatient services for diabetes using supervised methods. The average drug cost, average diagnostic fee, average amount claimed, average dispensing days, average medical expenditure per day, average counseling and treatment fees, average drug cost per day, average dispensing service fees, and average drug cost per day were the expenditure-related characteristics that they compared two groups of fraudulent and non-fraudulent (with/without fraud) claims to develop detection models. To detect fraud, three machine learning techniques—Classification Trees, Logistic Regression, and Neural Networks—were compared. All three approaches were successful in achieving accuracy although the Classification Tree model performed better overall, obtaining a 99% correct recognition rate.

Lin et al. (2008) used clustering techniques to analyze data about medical practitioners covered by the Taiwan National Health Insurance. A total of 10 attributes were used to categorize the physician data. The critical clusters were identified and ranked according to expert opinion regarding the influence of the clusters on health expenditures.

Aral et al. (2012) used a drug's commercial name, market price, prescription number, prescriber age, gender, and indication to detect prescription forgery. The model identified fraudulent prescriptions with a true positive rate of 77.4% and false positive rate of 6% in an adult cardiac surgery database. Similarly, Shin et al. (2012) examined 38 claims criteria to identify fraudulent claims in 3,705 outpatient internal medicine clinics. Using these features, a risk score was calculated, and a decision tree model was used to classify the providers. These studies demonstrate how prescription-specific data and broader claim characteristics can be used to detect and prevent medical fraud.

A model for detecting prescription forgeries was proposed by Aral et al. (2012). Six features were identified for this purpose: the prescribed drug's commercial name, the prescribed drug's market price, prescription number, the prescriber's age, gender, and the prescribed drug's indication. When evaluating the effectiveness of the model in identifying fraudulent medical prescriptions (with a true positive rate of 77.4% and a false positive rate of 6%) in the adult cardiac surgery database, the model demonstrated satisfactory performance in distinguishing between fraudulent and non-fraudulent prescriptions.

Shin et al. (2012) attempted to identify fraudulent claims in 3,705 outpatient internal medicine clinics. A total of 38 characteristics were identified from outpatient claims submitted to a health insurance organization as part of the research. Based on these attributes, a risk score was computed to represent the probability of fraudulent activity. Providers are classified using a decision tree model.

Srinivasan et al. (2013) used rule-based data mining to detect health insurance claim fraud, misuse, waste, and errors using Medicare data. This big data technique helps private health insurers find hidden cost overruns that standard information systems miss. In addition, Branting et al. (2016) developed a decision tree and graph analysis method to automatically detect and predict fraud. Through anomaly detection and predictive analysis, these methods provide a comprehensive framework to detect and prevent health insurance claim fraud.

A machine learning model was proposed by Bauder et al. (2016) to identify anomalous physician conduct in health insurance claims. The model attempts to identify instances in which physicians deviate from the established standards of their specialty, thereby notifying decision makers of billing procedure abuse or fraud. Through the use of five-fold cross-validation, the model sensitivity, specificity, and F1 score were computed. By employing the Naive Bayes algorithm, the model accurately forecasts multiple classifications of physicians with F1 scores exceeding 0.90.

Seven essential stages were outlined by Joudaki et al. (2015) as a framework for examining healthcare and insurance claims to detect fraud and abuse (following the data pre-processing phase). 1) Domain experts (Sokol et al., 2001; Li et al.) identify the most significant characteristics of the data; 2) Automated algorithms or expert opinion, such as

association analysis, identify new features that are indicative of fraudulent or malicious behavior (Li et al.); 3) Outlier detection methods identify unusual records for further investigation (Shan et al., 2009). 4) Process of extracting outliers from the dataset and clustering or clustering records according to the extracted features (Lin et al., 5) Determining outlier clusters and conducting additional analysis on the records contained within these clusters to detect fraudulent or malicious records 6) Constructing supervised models by selecting the most discriminatory features from the labeled records from the previous step (Liou et al., 2008). 7) Implementing supervised methods during regular online processing tasks and unsupervised methods (such as outlier detection and clustering) during designated time intervals in order to enhance preceding procedures and identify novel instances of fraudulent activity.

## 4. METHODOLOGY

### 4.1. Dataset<sup>1</sup>

The data used in this study were obtained from the "Healthcare Provider Fraud Detection Analysis" dataset on the Kaggle platform. This dataset is also a combination of the four datasets. The dataset contains retrospective data. These datasets are:

- Inpatient Data: This data contains information about the hospital claims submitted on behalf of inpatients. In addition, the admission and discharge dates, and the admission diagnostic code, are included.
- Outpatient Data: This data includes claim details for patients who visited the hospitals but were not admitted.
- Beneficiary Details Data: This data consists of information about the beneficiaries, including their ages, health conditions, region, and more.
- Target Class (PotentialFraud): This class includes fraud classes. (Yes/No).

This dataset was acquired from the Kaggle platform's "Healthcare Fraud Detection Analysis" dataset, is also composed of four datasets. The final dataset was acquired via data preprocessing and feature engineering. The final dataset and its descriptions are presented in Table 1.

### 4.2. Data Preprocessing/Feature Engineering

In the first version of the dataset, before conducting a comprehensive analysis of the dataset and prior to conducting the exploratory data analysis phases, 4904 non-fraudulent and 506 fraudulent activities were observed (Table 1). The "BeneID" feature, which is shared by all four datasets, was used to merge the datasets. This feature comprises individualized patient identities. There were 212796 fraudulent records (fraudulent-Yes) and 345415 normal records (non-fraudulent-No) in the final stage. During the phase of exploratory data analysis, "pandas-profiling" module of the Python library pandas was implemented. By using the extremely high-level application programming interface (API) provided by pandas profiling, a data scientist can generate an exhaustive profile report. Five major sections comprise the output report: Introduction, Variables, Interactions, Correlation, and Missing Values. Pandas-profiling is widely regarded as the most powerful Python library for exploratory data analysis (Brugman, 2019).

It was detected in the output report that certain attributes (DOD and CLIMPROCEDURECode) were missing values. The dataset was eliminated of these characteristics. Consequently, 133980 outpatient and 31289 inpatient records were acquired. At this point in the dataset, three distinct physician types (specialist, operator, and other) were merged into a single physician, for 100737 distinct physicians. The most crucial stage in comprehending the data and, consequently, detecting fraud consists of answering inquiries such as the number of beneficiaries in the dataset, the quantity and variety of physicians, the standing of physicians affiliated with various providers/hospitals, the correlation between patient age and medical claims, the correlation between patient age and chronic conditions, and so forth.

The goal of feature engineering is to select effective feature sets for different fraud models. After feature engineering, the dataset shown in Table 1 was obtained. Table 1 provides a description of each feature in the dataset, a description of the feature, and statistical information (range- max/min values-distribution-0% missing values) for each feature.

<sup>1</sup> <https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis>



**Table 1.** Dataset used for modeling

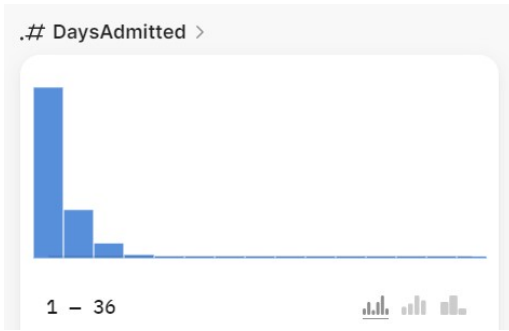
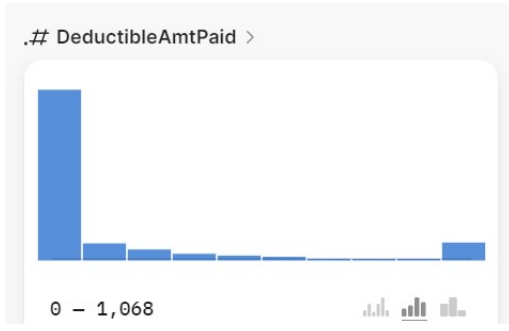
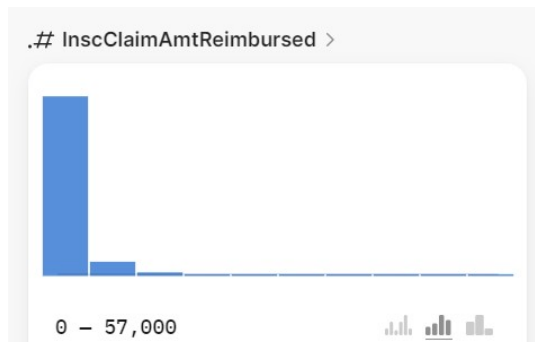
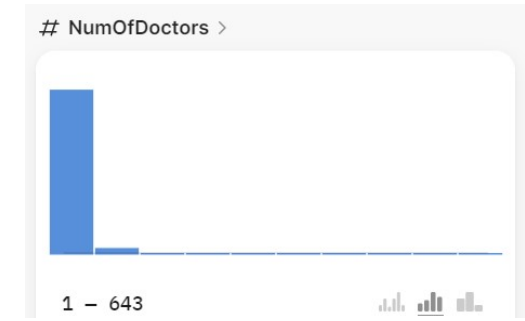
Feature	Description	Statistical Information
<b>DaysAdmitted</b>	Duration of the inpatient-hospital stay (days)	 <p>.# DaysAdmitted &gt;</p> <p>1 - 36</p>
<b>DeductibleAmtPaid</b>	Amount paid by the patient. (total amount requested-amount reimbursed)	 <p>.# DeductibleAmtPaid &gt;</p> <p>0 - 1,068</p>
<b>Provider</b>	Provider's ID	
<b>InscClaimAmtReimbursed</b>	Amount reimbursed for the claim	 <p>.# InscClaimAmtReimbursed &gt;</p> <p>0 - 57,000</p>
<b>NumOfDoctors</b>	Number of physicians	 <p># NumOfDoctors &gt;</p> <p>1 - 643</p>

Table 1. Continued

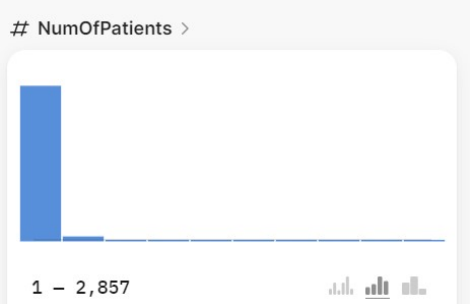
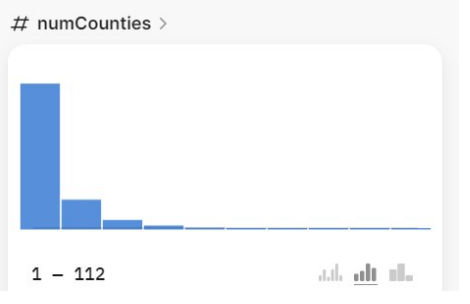
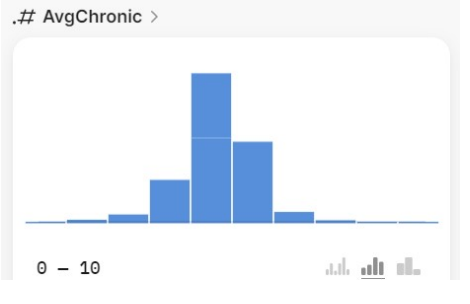




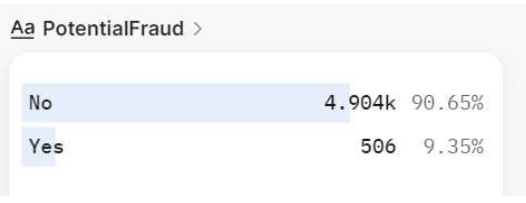
<p><b>NumOfPatients</b></p>	<p>Number of patients</p>	
<p><b>numCountries</b></p>	<p>Number of countries (for each provider)</p>	
<p><b>AvgChronic</b></p>	<p>Average chronic diseases</p>	
<p><b>AvgClaim</b></p>	<p>Average claim</p>	
<p><b>Age</b></p>	<p>Patient's age</p>	

Table 1. Continued

<b>male</b>	Number of men	
<b>female</b>	Number of women	
<b>PotentialFraud</b>	Whether the recording is fraudulent (target class)	

### 4.3. Modeling

Because the dataset is an imbalanced dataset; in other words, non-fraudulent records are more than fraudulent records, XGBoost, LGBM, and Random Forest (RF), which are ensemble learning-based models that work well on imbalanced datasets, were preferred. In addition, Logistic Regression (LR) was used because the binary classification problem was addressed after the data were balanced using the SMOTE method. The hyperparameters were first trained using the default parameters provided by the model library in Python. For the RF model, hyperparameter optimization was also performed (max depth: 4, number of trees: 500). In the modeling phase, the data were divided into a 70% training set and a 30% test set. The machine learning models employed are described in Table 2.

## 5. FINDINGS

In this section, the results of the LR, NB, SVM, RF, XGBoost, and LightGBM models applied to the preprocessed dataset are evaluated according to different performance metrics. The applied model parameters are first trained on the default parameters. As the accuracy metric may be misleading due to imbalances in the dataset, sensitivity/recall, and precision metrics are included.

The ROC/AUC metric for the LR model was also included in the findings. The confusion matrix, a matrix summarizing the performance of a machine learning model on a set of test data, is also discussed. Table 3 lists the performance metrics of the models. When evaluating classification models that attempt to predict a categorical label for each input sample, a confusion matrix is frequently used.

The number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) generated by the model on the test data is displayed in the matrix. Figure 1 shows the confusion matrix resulting from the RF model. The ROC Curve for the LR model is shown in Figure 2.

**Table 2.** Machine Learning Models

<b>Model</b>	<b>Description</b>
<b>Logistic Regression (LR)</b>	When using logistic regression, it is important to consider the connections between discrete variables. Logistic regression differs significantly from linear regression in that it uses binary or multiple outcome variables rather than numeric variables. As Bircan (2010) pointed out, this technology has widespread application in healthcare.
<b>Naive Bayes (NB)</b>	The Naive Bayes algorithm is a supervised machine learning algorithm that relies on conditional probability and applies the Bayes' theorem (Vangara et al., 2020). The proposed approach relies on the assumption that the properties of the input data are conditionally independent given class, which allows the algorithm to make quick and precise predictions.
<b>Random Forest (RF)</b>	An ensemble learning classifier called Random Forest (RF) generates multiple decision trees by selecting utilizing a randomly selected subset of training samples and variables. The RF classifier uses a CART collection to generate predictions (Breiman, 2001). Trees are generated by selecting a subset of training examples using a replacement method known as the bagging approach. It is possible for certain samples to be selected multiple times, whereas others may not be selected at all (Belgiu and Draĝut, 2016: 25).
<b>Support Vector Machine (SVM)</b>	Classification was performed by SVM using either a linear or nonlinear function. This involves estimating the most suitable function to separate the data, as Özkan (2016) explained. The algorithm places all feature vectors in a virtual space and divides the samples by a line known as a hyperplane. The hyperplane was designed to effectively separate classes by maximizing the margin (Burkov, 2019).
<b>XGBoost</b>	The development of this method was performed by Chen et al. (2016). This implementation of gradient boosting machines is highly advanced and can greatly enhance the computational power of boosting tree algorithms. It was created with a strong focus on optimizing model performance and computational speed. Boosting is a powerful technique in ensemble learning that involves adding new models to correct errors caused by existing models. Models are added to the model recursive until any further improvement is no longer detected (Ogunleye and Wang, 2020: 2133).
<b>LightGBM</b>	Ever since its introduction by Ke (2017), LightGBM has attracted significant research attention. LightGBM is a highly efficient implementation of gradient-boosting trees, and it is known for its adaptability and effectiveness. LightGBM primarily uses histogram algorithms and other algorithms to enhance the computational power and prediction accuracy of the algorithm. First, continuous feature values are expressed in M integers, and then a histogram of width M is decomposed. Using the decomposed values of the histogram, the data is analyzed to determine the decision tree. By leveraging the histogram algorithm, significant improvements can be made to the time complexity. Additionally, the fuzzy partitioning method outperforms the decision tree, making it a valuable approach (Wang et al., 2022: 261).

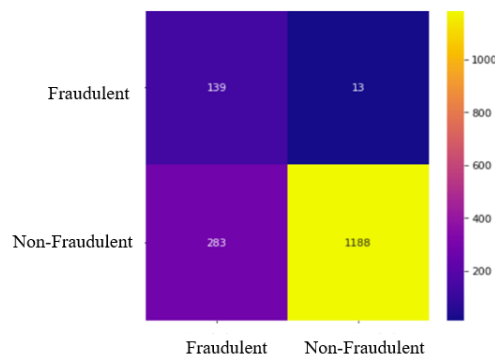
**Table 3.** Model Performances

<b>Model</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Sensitivity/ Recall</b>	<b>F1-Score</b>
<b>LR</b>	0.820	0.796	0.823	0.809
<b>NB</b>	0.896	0.94	0.95	0.945
<b>SVM</b>	0.918	0.92	1.00	0.958
<b>RF</b>	0.823	0.808	0.914	0.858
<b>XGBoost</b>	0.925	0.95	0.97	0.96
<b>LightGBM</b>	0.930	0.95	0.98	0.965

There is a trade-off between Precision and Recall metrics, especially in healthcare. For example, high precision requires low FP; thus, a classifier that maximizes precision only returns strongly positive predictions, which may result in missing positive occurrences. In this context, which of the precision and recall metrics are maximized depends on the application. Recall is a statistic used when the expense of estimating FN is significant. It should be maximized. For example, if a Fraud Detection model misclassifies a fraudulent transaction as non-fraudulent, it may have adverse implications for the bank. Therefore, the Recall value is of critical importance because of the performed analyses.

As shown in Table 3, the SVM model performed the best. This is not surprising in the context of the SVM model structure. Fraud detection frequently involves complicated sets of data with numerous features. SVM is particularly suitable for datasets with many dimensions because it can effectively manage a large number of input parameters without observing a noticeable decline in performance. This is especially advantageous when handling sophisticated transactional data that involve multiple features.

The SVM model is followed by ensemble learning models (LightGBM, XGBoost, RF) when the Recall value is considered. When the precision value was considered, it was observed that the model with the best performance was LightGBM. The reasons for the high performance of ensemble learning models are that they require minimal data preprocessing in classification problems, are not sensitive to outliers, and can work with imbalanced and missing data, as well as the acquired dataset. As mentioned previously, the LightGBM and SVM models performed best on the recall metric. Figures 1 and 2 were chosen to illustrate the different models and their evaluation methods. Therefore, confusion matrix and ROC curve, which are the most commonly used modeling methods to provide visual support, were preferred.

**Figure 1.** Confusion Matrix (RF)

In Figure 1, the confusion matrix for RF can be observed in terms of fraudulent and non-fraudulent activities. As can

be seen, The Random Forest model correctly predicted 139 of 152 fraudulent activities and 1188 of 1471 non-fraudulent activities.

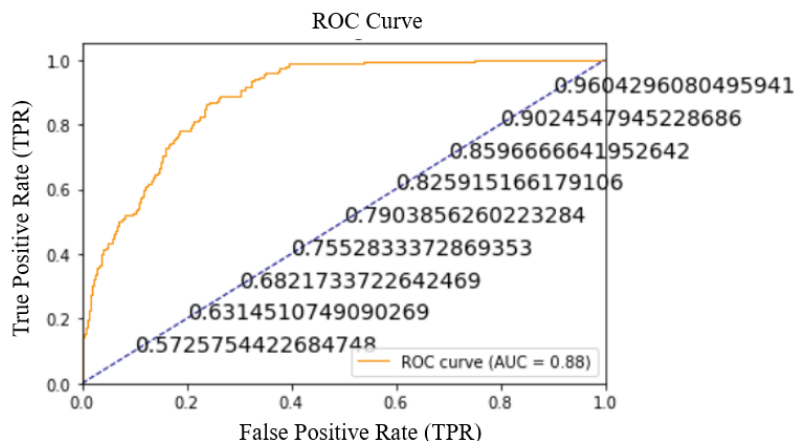


Figure 2. ROC Curve (LR)

Figure 2 shows the ROC curve for logistic regression. This method, which is frequently used, especially for classification problems in the medical domain, graphically presents the relationship between sensitivity (TP) and 1-specificity (FP) values to determine an ideal “cut-off” value. The area under the curve (AUC) provides a good measure of discrimination (Fan et al., 2006: 19). Figure 2 shows the ROC space. Accordingly, a well-performing classifier is expected to be near the upper left corner of the graph. As can be seen from the ROC curve for the LR model, the classifier performance reached approximately 96% because of the analysis. This demonstrates the success of the proposed LR model in binary classification.

## 6. DISCUSSION AND CONCLUSION

Medical fraud is a significant concern for numerous health systems. It occurs when an individual deliberately (willingly) submits false or deceitful statements, orchestrates the dissemination of such statements, or manipulates the facts to secure payment for healthcare services to which they are not legally entitled. Medical fraud may manifest itself through various means, including yet not limited to referrals for government-prohibited health care services, bribery, or kickbacks for publicly funded health care services (Centers for Medicare & Medicaid Services, 2021:6). With greater precision, the occurrences of medical fraud in the healthcare industry vary extensively and encompass a wide array of techniques and procedures. Schemes to commit medical fraud may involve the work of a single individual or the collaboration of an organization. Infiltrating healthcare systems’ programs and functioning as healthcare providers or suppliers is not beyond the capability of organized criminal syndicates (Centers for Medicare & Medicaid Services, 2021:6).

The exact cost of medical fraud in society is unknown. According to estimates from the National Health Care Anti-Fraud Association, Medicare and Medicaid fraud cost taxpayers more than 100 billion a year (Zamost and Brewer, 2023). According to another estimate, the Medicare program in the US provides healthcare to more than 60 million US citizens, whereas Medicare loses \$20 to \$70 billion annually due to fraud, waste, and abuse (Johnson and Khoshgoftaar, 2019:31). Over the projected period, the worldwide healthcare fraud market is anticipated to increase at an annual growth rate of 20.45%, from US\$1.65 billion in 2022 to US\$5.03 billion in 2028 (Arizton, 2022).

Medical fraud can put the health and well-being of healthcare beneficiaries at risk and cost billions of dollars to many sectors of society, including beneficiaries of healthcare programs, healthcare institutions and organizations, and taxpayers. As countries’ health services continue to serve increasing numbers of beneficiaries, the impact of these losses and risks grows. Today, medical fraud can frequently be detected and avoided using machine learning technologies that have an immense impact on health programs’ capability to offer affordable, high-quality medical services.

This paper presents a comprehensive performance analysis of an example of the implementation of machine learning algorithms for medical fraud detection. This study uses machine learning methods to binary classify cases into cases where fraud is present and cases where fraud is absent to detect and correct fraud with higher accuracy. More precisely,

automatic classification of medical fraud can be achieved using machine learning. In the analysis phase, LR, Naïve Bayes, SVM, RF, XGBoost, and LightGBM models, which are frequently used for binary classification, were used. In evaluating the results of the analysis, not only the accuracy, precision and recall metrics were used to the imbalanced nature of the dataset. When focusing on accuracy and precision, the LightGBM model appears to be the best performing model (90.3%); however, in terms of precision, the SVM model yielded the highest results. In this case, the preferred model will vary depending on whether only the recall ratio is taken into account or the F1 score, which is the harmonic mean of the precision and recall ratio, is calculated and included in the analysis.

The potential for medical fraud to be detected by AI systems may raise the anxiety of healthcare professionals, including medical secretaries, billing specialists, and auditors, regarding their employment prospects. The potential outcome of AI taking over these tasks is that employees may experience anxiety regarding their employment status. Öztürk (2023) and Özbek (2024) examined the AI concerns of accounting professionals and innovation-oriented behaviors of employees, respectively. Nevertheless, this study primarily focused on the collaboration between humans and artificial intelligence. Simply put, when AI identifies fraudulent activities in extensive data collections, employees can assist the AI by examining and confirming these findings. This collaboration can help ensure the continued significance of employee responsibilities and foster a positive perception of AI as a supportive tool rather than a source of concern.

The demonstration of the significant performance of machine learning over other approaches, especially manual detection methods, proves that machine learning algorithms will be used much more frequently in future fraud detection applications.

---

**Peer Review:** Externally peer-reviewed.

**Author Contributions:** Conception/Design of Study- C.Ü., G.S.E.; Data Acquisition- C.Ü., G.S.E.; Data Analysis/Interpretation- C.Ü., G.S.E.; Drafting Manuscript- C.Ü., G.S.E.; Critical Revision of Manuscript- C.Ü., G.S.E.; Final Approval and Accountability- C.Ü., G.S.E.

**Conflict of Interest:** The authors have no conflict of interest to declare.

**Grant Support:** The authors declared that this study has received no financial support.

---

### ORCID IDs of the authors

Ceyda ÜNAL 0000-0002-5503-8124  
Gökçe Sinem ERBUĞA 0000-0003-1604-4668

### REFERENCES

- Abdallah, A., Maarof, M.A. and Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68, 90-113.,
- Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.
- Altındış, S., & Morkoç, İ. K. (2018). Sağlık hizmetlerinde büyük veri. Nigde Omer Halisdemir University *Academic Review of Economics and Administrative Sciences*. 11(2), 257-271.
- Aral, K. D., Güvenir, H. A., Sabuncuoğlu, İ., & Akar, A. R. (2012). A prescription fraud detection model. *Computer Methods and Programs in Biomedicine*, 106(1),37-46. <http://dx.doi.org/10.1016/j.cmpb.2011.09.003>
- Arizton. (2022). Healthcare Fraud Analytics Market- Global Outlook & Forecast 2023-2028. <https://www.arizton.com/market-reports/healthcare-fraud-analytics-market>
- Avcı, M., and Teyyare, E. (2012). Sağlık sektöründe yolsuzluk: Teorik bir değerlendirme. *TheInternationalJournalofEconomicandSocialResearch(IJESR)* 8(2). 199-221.
- Aydın, J. C., & Yaşar, G. (2020). Sağlık Harcamalarının Gelir Esnekliği Açısından Değerlendirilmesi: Sistematik Bir Derleme. *JournalofAnkaraHealthSciences*,9(1), 63-80.
- Aydoğan Duman & Sağıroğlu, Ş. (2017, October). Health care fraud detection methods and new approaches. *2017 International Conference on Computer Science and Engineering (UBMK)* (pp. 839-844). IEEE.
- Bauder, R. A., and Khoshgoftaar, T. M. (2016, December). A probabilistic programming approach for outlier detection in healthcare claims. *2016 15th IEEE international conference on machine learning and applications (ICMLA)* (pp. 347-354). IEEE.
- Bauder, R. A., Khoshgoftaar, T. M., Richter, A., & Herland, M. (2016, Kasım). Predicting medical provider specialties to detect anomalous insurance claims. *2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI)* (ss. 784-790). IEEE.
- Belgiu, M., & Drăguț, L. (2016). Random Forest in Remote Sensing: A review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24-31.

- Bircan, H. (2004). Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. *Kocaeli University Journal of Social Sciences*, (8), 185-208.
- Branting, L. K., Reeder, F., Gold, J., & Champney, T. (2016, Ağustos). Graph analytics for healthcare fraud risk estimation. *2016 IEEE/ACM Intl Conf. on Advances in Social Networks Analysis and Mining (ASONAM)* (ss. 845-851). IEEE.
- Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit scoring and credit control VII*, 235-255.
- Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, 17(3), 235-255.
- Bozhenko, V. (2022). Tackling corruption in the health sector. *Health Economics and Management Review*, 3, 32-39.
- Brugman, S. (2019). pandas-profiling: exploratory data analysis for Python. <https://github.com/pandas-profiling/pandas-profiling>.
- Burkov, A. (2019). Hundred-page Machine Learning Book. Quebec City, QC, Canada: Andriy Burkov.
- Centers for Medicare and Medicaid Services (CMS). (2021). Medicare Fraud & Abuse: Prevent, Detect, Report. <https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/Fraud-Abuse-MLN4649244.pdf>
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACCM signed international conference on knowledge discovery and data mining* (ss. 785-794).
- Couffinhal, A., and Frankowski, A. (2017). Wasting with intention: Fraud, abuse, corruption, and other integrity violations in the health sector.
- Dora, P. and Sekharan, G. H. (2015). Healthcare insurance fraud detection leveraging big data analytics. *IJSR*, 4(4), 2073-2076.
- Ekin, T. (2019). An Integrated Decision-Making Framework for Medical Audit Sampling.
- Ekin, T., Frigau, L., & Conversano, C. (2021). Health care fraud classifiers in practice. *Applied Stochastic Models in Business and Industry*, 37(6), 1182-1199.
- European Commission, Directorate-General for Migration and Home Affairs (EC). (2017). Weistra, K., Swart, L., Oortwijn, W., et al., *Updated study on corruption in the healthcare sector: final report*, Publications Office. <https://op.europa.eu/en/publication-detail/-/publication/9537ddb7-a41e-11e7-9ca9-01aa75ed71a1/language-en>
- Euronews. (2022). Sağlık harcamalarının milli gelire oranı: OECD ve AB’de sağlığa en çok pay ayıran ülkeler hangileri? Retrieved from <https://tr.euronews.com/next/2022/04/05/sagl-k-harcamalar-n-n-milli-gelire-oran-oecd-ve-ab-de-sagl-ga-en-az-pay-ay-ran-ulke-turkiy>
- J. Fan, S. Upadhye. Ve Worster, A. (2006). Understanding Receiver Operating Characteristic (ROC) Curves. *Canadian Journal of Emergency Medicine*, 8(1), 19-20.
- Federal Bureau of Investigation (FBI). (1989). White collar crime: a report to the public. Washington, DC: Government Printing Office.
- Francis, C., Pepper, N., & Strong, H. (2011, Ağustos). Using support vector machines to detect medical fraud and abuse. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (ss. 8291-8294). IEEE.
- Gee, J., Button, M., & Brooks, G. (2010). The financial cost of healthcare fraud: data from around the world. (London: MacIntyre Hudson/CCFS).
- Georgetown University Memory Disorders Program (2023). Allegations of fraud in Alzheimer’s disease research: Death of the amyloid hypothesis?, <https://memory.georgetown.edu/news/allegations-of-fraud-in-alzheimers-disease-research-death-of-the-amyloid-hypothesis%EF%BF%BC/>
- He, H., Graco, W., & Yao, X. (1999). Application Of Genetic Algorithm And K-Nearest Neighbor Method İn Medical Fraud Detection. *Lecture Notes in Comput. Sci.* 1585 74–81. Springer, Berlin.
- He, H. X., Wang, J. C., Graco, W., & Hawkins, S. (1997). Application of Neural Networks detect Medical Fraud. *Expert Systems with Applications* 13 329–336.
- ISSA. (2022). Detecting fraud in healthcare through emerging Technologies. <https://ww1.issa.int/analysis/detecting-fraud-health-care-through-emerging-technologies>
- Johnson, J. M., & Khoshgoftaar, T. M. (2019). Medicare fraud detection using neural networks. *Journal of Big Data*, 6(1), 1-35.
- Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., and Arab, M. (2015). Using data mining to detect healthcare fraud and abuse: a review of literature. *Global Journal of Health Sciences* 7(1), 194.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma W, et al. & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- Kirlidog, M., and Asuk, C. (2012). A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences*, 62, 989-994.
- Kurşun, A. (2021). Büyük Veri ve Sağlık Hizmetlerinde Büyük Veri İşleme Araçları. *Hacettepe Sağlık İdaresi Dergisi*, 24(4), 921-940.
- Küçük, A. (2022). Sağlık Hizmet Ödemelerinde Usulsüzlük Türleri ve Mücadele Stratejileri. *Sayıştay Dergisi*, 33(127), 585-607.
- Li, J., Huang, K. Y., Jin, J., & Shi, J. (2008). Survey on statistical methods for healthcare fraud detection. *Health Care Management Science*, 11(3), 275-287. <http://dx.doi.org/10.1007/s10729-007-9045-4>.
- C. Lin, C. M. Lin, S. T. Li, and S. C. (2008). Intelligent physician segmentation and management based on KDD approach. *Expert Systems with Applications*, 34(3), 1963-1973. <http://dx.doi.org/10.1016/j.eswa.2007.02.038>
- Liou, F. M., Tang, Y. C., & Chen, J. Y. (2008). Detecting hospital fraud and claim abuse through diabetic outpatient services. *Health Care Manag Sci. II*, 353-358.
- Liu, Q., & Vasarhelyi, M. (2013, Kasım). Healthcare fraud detection: A survey and clustering model incorporating geo-location information. *Brisbane, 29th World Continuous Auditing and Reporting Symposium (29WCARS)*.
- Major, J. A., & Riedinger, D. R. (1992). Efd: A Hybrid Knowledge/Statistical-Based System for Detecting Fraud. *International Journal Of Intelligent Systems*. 687–703.
- National Health Care Anti-Fraud Association (NHCAA). (2021). The Challenge of Health Care Fraud. <https://www.nhcaa.org/tools-insights/>



[about-health-care-fraud/the-challenge-of-health-care-fraud/](#)

- OECD (2020). Public Integrity for an Effective COVID-19 Response and Recovery. [https://read.oecd-ilibrary.org/view/?ref=129\\_129931-ygq2xb8qax&title=PublicIntegrityforanEffectiveCOVID-19ResponseandRecovery](https://read.oecd-ilibrary.org/view/?ref=129_129931-ygq2xb8qax&title=PublicIntegrityforanEffectiveCOVID-19ResponseandRecovery)
- Ogunbanjo, G. A., and D. K. (2014). Ethics in health care: healthcare fraud. *South African Family Practice*, 56(1), S10-S13.
- Ogunleye, A., & Wang, Q. G. (2019). XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6), 2131-2140.
- Orhan, M. S., & Serçemeli, M. (2015). İç denetim stratejisinde sürekli denetimin uygulanabilirliğine ilişkin bir araştırma. *Giresun Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 1(2), 83-110.
- Özbek, A. (2024). Muhasebe Meslek Mensuplarının Yapay Zekâ Kaygılarının Gelecekte İstihdam Edilebilirlik Algıları Üzerine Bir Çalışma. *Alanya Akademik Bakış*, 8(1), 254-267.
- Özkan, Y. (2016). Veri Madenciliği Yöntemleri. İstanbul: Papatya Yayıncılık.
- Öztürk, M. (2023). A study on the impact of artificial intelligence anxiety on innovation-oriented behaviors of employees. *Optimum Ekonomi ve Yönetim Bilimleri Dergisi*, 10(2), 267-286.
- Price, M., & Norris, D. M. (2009). Health care fraud: physicians as white-collar criminals? *Journal of the American Academy of Psychiatry and the Law Online*, 37(3), 286-289.
- Saravanan, R., and Sujatha, P. (2018, Haziran). State-of-the-art machine learning algorithms: a perspective of supervised learning approaches in data classification. In *2018 Second international conference on intelligent computing and control systems (ICICCS)* (ss. 945-949). IEEE.
- Savedoff, W. D., & Hussmann, K. (2006). The causes of corruption in the healthcare sector: a focus on health care systems. *Transparency International. Global Corruption Report*, 4-16.
- Shan, Y., Jeacocke, D., Murray, D. W., & Sutinen, A. (2008, Kasım). Medical specialist billing patterns for health service management. *Proceedings of the 7th Australasian Data Mining Conference*, 87 (ss. 105-110).
- Shan, Y., Murray, D. W., & Sutinen, A. (2009, Aralık). Discover inappropriate billings using a local density-based outlier detection method. *Proceedings of the Eighth Australasian Data Mining Conference*, 101 (ss. 93-98).
- Shin, H., Park, H., Lee, J., & Jhee, W. C. (2012). A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications*, 39(8), 7441-7450. <http://dx.doi.org/10.1016/j.eswa.2012.01.105>.
- Sokol, L., Garcia, B., Rodriguez, J., West, M., & Johnson, K. (2001). Using data mining to find fraudulent HCFA healthcare claims. *Topics in Health Information Management*, 22(1), 1-13.
- Srinivasan, U., & Arunasalam, B. (2013). Leveraging big data analytics to reduce healthcare costs. *IT professional*, 15(6), 21-28.
- Thomson Reuters. (2021). Organized crime using sophisticated technology in next wave of government healthcare fraud schemes. <https://www.thomsonreuters.com/en-us/posts/investigation-fraud-and-risk/healthcare-fraud-webinar/>
- Transparency International. (2006). Global Corruption Report 2006. [https://images.transparencycdn.org/images/2006\\_GCR\\_HealthSector\\_EN.pdf](https://images.transparencycdn.org/images/2006_GCR_HealthSector_EN.pdf)
- Transparency International. (2023). Health. <https://www.transparency.org/en/our-priorities/health-and-corruption>
- Turgay, İ., Doğan, S., & Mengi, B. T. (2020). İç Denetim Faaliyetlerinde Sürekli Denetim: Analitik İnceleme Prosedürlerinin Kullanımı. *Denetim*, (21), 5-26.
- U.S. Department of Justice (2023). Retrieved from <https://www.justice.gov/opa/pr/hospice-medical-director-sentenced-150m-hospice-fraud-scheme>.
- U.S. Department of Justice (2024). Criminal Resource Manual CRM 500-999, 976- Health Care Fraud—Generally. <https://www.justice.gov/archives/jm/criminal-resource-manual-976-health-care-fraud-generally#>
- USSC (2022). Quick Facts—Health Care Fraud Offenses. [https://www.uscc.gov/sites/default/files/pdf/research-and-publications/quick-facts/Health\\_Care\\_Fraud\\_FY22.pdf](https://www.uscc.gov/sites/default/files/pdf/research-and-publications/quick-facts/Health_Care_Fraud_FY22.pdf)
- Van Capelleveen, G., Poel, M., Mueller, R. M., Thornton, D., and van Hillegersberg, J. (2016). Outlier detection in healthcare fraud: A case study in the Medicaid dental domain. *International Journal of Accounting Information Systems* 21, 18-31.
- Vangara, V., Vangara, S. P., & Thirupathur, K. (2020). Opinion mining classification using naive bayes algorithm. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(5), 495-498.
- Vian, T. (2020). Corruption and administration in healthcare. In *Handbook on corruption, ethics, and integrity in public administration* (pp. 115-128). Edward Elgar Publishing.
- Vincke, P. ve Cylus, J. (2011). Healthcare fraud and corruption in Europe: An overview. *Eurohealth*, 17(4), 14-18.
- Wang, D. N., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. *Information Sciences*, 602, 259-268.
- World Health Organization (WHO). (2023a). *World health statistics 2023: monitoring health for the SDGs, sustainable development goals*. World Health Organization. Retrieved from <https://www.who.int/publications-detail-redirect/9789240074323>.
- World Health Organization (WHO). (2023b). Reducing Health System Corruption. <https://www.who.int/activities/reducing-health-system-corruption>
- Zamost, S. & Brewer, C. (2023). Inside the mind of criminals: How to brazenly steal \$100 billion from Medicare and Medicaid. <https://www.cnbc.com/2023/03/09/how-medicare-and-medicaid-fraud-became-a-100b-problem-for-the-us.html#:~:text=Fraud%20flourishes&text=Taxpayers%20are%20losing%20more%20than,Health%20Care%20Anti%2DFraud%20Association>.
- Zhang, C., Xiao, X., & Wu, C. (2020). Medical fraud and abuse detection system based on machine learning. *International journal of environmental research and public health*, 17(19), 7265.

Zhu, S., Wang, Y., & Wu, Y. (2011, Ađustos). Health care fraud detection using nonnegative matrix factorization. 2011 6th International Conference on Computer Science & Education (ICCSE) (ss. 499-503). IEEE.

### **How cite this article**

Ůnal, C., & Erbuža, G. S. (2024). Detection and Prevention of Medical Fraud using Machine Learning. *Acta Infologica*, 8(2), 100-117. <https://doi.org/10.26650/acin.1463879>