Computing and Informatics, Vol. 45, 2026, 800–829, doi: 10.31577/cai_2026_4_800

ACTING LIKE OR THINKING LIKE HUMAN: DOCTORS AND DEEP LEARNING ALGORITHMS BEHAVIORAL COMPARISON

Yusuf Altuntaş

Department of Orthopedics and Traumatology Service
Şişli Hamidiye Etfal Research and Training Hospital
İstanbul, Türkiye
e-mail: yusuf.altuntas1@saglik.gov.tr

Muhammed Taha Zeren

Merkezi Kayıt Kuruluşu A.Ş.
Central Securities Depository and Trade Repository of Türkiye
İstanbul, Türkiye
e-mail: muhammed.zeren@mkk.com.tr

Şebnem Özdemir

Management Information System Department
İstinye University
İstanbul, Türkiye
e-mail: sebnem.ozdemir@istinye.edu.tr

Abstract. This study aims to compare the behavioral outputs and response durations' analysis of two distinct groups of medical professionals – general practitioners and orthopedic specialist doctors – utilizing deep learning and computer vision algorithms. This comparison is facilitated through one-way ANOVA and Tukey pairwise comparison tests. The research focuses on femoral proximal fractures, which have the highest disability and mortality rates among the elderly population globally. For this research, over 1 500 femoral proximal region images sourced from

over 500 patients aged between 22 and 105 for retraining of deep learning algorithms. Updated SSD – MobileNet v2, Updated Faster R-CNN – Inception v2 and Updated YOLO Darknet v4 algorithms were retrained specifically for femoral proximal fracture detection. In this study, a total of 50 X-ray images comprising femoral proximal fractures and a total of 50 X-ray images without fractures were gathered, in total evaluation dataset containing 55 X-ray images, underwent scrutiny via one-way ANOVA tests, assessing both tree retrained computer vision deep learning algorithms, as well as physician groups concerning fracture detection success in fractured and non-fractured regions, for the entirety of the evaluation dataset. Subsequent Tukey pairwise comparison tests were conducted based on the obtained results. Regarding result production durations, it was observed that the average detection time of expert physicians significantly exceeded that of general practitioners, the Updated YOLO algorithm, the Updated Faster R-CNN algorithm, and Updated SSD Algorithm. Conversely, no significant disparity was noted between the average result production times of general practitioners and expert physicians. In Tukey pairwise tests evaluating the accuracy rate of fractured regions, the success rate of expert physicians in fractured region accuracy significantly surpassed that of SSD and general practitioners, while insignificantly differing from Faster R-CNN and YOLO Algorithms. For non-fractured region accuracy, the Tukey pairwise test revealed that the accuracy rate of the YOLO algorithm significantly outstripped that of Faster R-CNN and SSD Algorithms, while insignificantly differing from expert and general practitioners. However, it was noted that the success rate of Faster R-CNN and SSD algorithms significantly differed from expert physicians. In conclusion, individual output behavior differs between general practitioners and expert physicians, but expert physicians show (like AI Algorithms) similar behaviors in the same outputs.

Keywords: Computer vision, deep learning, machine learning, artificial intelligence, artificial neural networks, YOLO, Faster R-CNN, SSD, ANOVA, Tukey pairwise test

1 INTRODUCTION

Humanity stands on the brink of a turning point that has never been experienced in known human history. It is impossible not to witness this while every second of life flows by, or even while performing any daily task. Although more than half a century has passed since the questions "Can mechanical systems think?" or "Can the characteristics of human intelligence be observed in a system?" were posed, it is known that the existence of artificial intelligence and the questions concerning this existence date back much further [1, 2]. One of the most significant pieces of evidence of this is the autonomous mobility-enabled chairs, referred to as "tripods", described in Homer's Iliad, in the 7^{th} and 8^{th} centuries BC [3]. In subsequent centuries, including Aristotle, who lived between the 4^{th} and 3^{rd} cen-

turies BC, both Greek thinkers and those from other regions of the world debated whether inanimate objects could possess autonomous capabilities or the ability to think.

Since its designation as "Artificial Intelligence" in 1956 and its attainment of its current competence, artificial intelligence has experienced two periods of dormancy, known as "AI winters", due to various reasons, primarily financial and technical challenges. During these periods, when technological and technical capabilities of processors were insufficient, the agenda shifted, and faith in artificial intelligence was lost. However, with the development of computer and processor technologies that accelerated in the 1980s and the subsequent advances in algorithms and artificial neural networks in the early 1990s, artificial intelligence not only awakened from what was considered an endless winter sleep but also resurged for victories against human intelligence [4, 5, 6, 7, 8].

In 2019, one of the most debated topics at the World Artificial Intelligence Conference (WAIC) in Shanghai, attended by owners of leading global technology firms, was whether artificial intelligence could surpass human intelligence. An important argument put forth in these discussions was that artificial intelligence, being the creation and guardian of human intelligence, could not surpass it. However, developing technology and real-life cases have shown the contrary. One of the earliest scenarios in which smart systems and human intelligence competed as rivals was in 1997, when Garry Kasparov faced off against the intelligent system known as Deep Blue. Although it is still debated whether there was human intervention that enabled the AI to win, IBM's chess computer, Deep Blue, made history by defeating the world chess champion Garry Kasparov, albeit controversially [9]. In 2011, IBM's Watson, an AI developed by IBM, defeated two former champions in the "Jeopardy!" quiz show using its natural language processing and data mining capabilities [10]. In 2016, Google DeepMind's AlphaGo program took on Lee Sedol, the world champion of the much more complex game of Go than chess. Without any controversy, AlphaGo defeated the world champion Lee Sedol with a decisive 4–1 victory [11].

Developed by Carnegie Mellon University, an artificial intelligence named Libratus defeated four professional poker players in a game that included human characteristics such as information gaps and bluffing [12]. By 2019, in the dynamic and rapidly changing strategy game Dota 2, OpenAI Five, an intelligent system, defeated a team of professional players [13]. In 2019, IBM's Project Debater AI participated in a debate competition against humans. This developed AI demonstrated its ability to debate against humans, utilizing its knowledge synthesis and natural language processing capabilities [14]. In 2020, OpenAI's GPT-3 model showed performance close to humans in language processing and creative writing, demonstrating the potential of AI in creative thinking and content generation [15]. The DeepMind team's AI named AlphaFold revolutionized biology and medicine by solving the protein folding problem in 2020. This achievement has shown how AI can support humans in science [16]. In 2024, Zheng and colleagues discussed ChatGPT as a potentially useful supplementary resource in the management of rare

and complex diseases. Its language-processing capabilities may assist clinicians in interpreting symptom descriptions, recognizing conditions that warrant further investigation, and retrieving relevant information from extensive medical literature. It may also support the evaluation of patient-specific treatment options by integrating medical history, clinical data, and existing clinical guidelines. However, these potential applications do not position ChatGPT as an independent diagnostic tool. Its outputs may contain inaccurate, incomplete, or biased information and therefore should be carefully reviewed by qualified healthcare professionals before being applied in clinical decision-making [17].

In 2023, researchers at the University of California, San Francisco developed an AI-supported tool to diagnose acute hepatic porphyria (AHP), a rare genetic disorder. By analyzing electronic health records, this algorithm was able to identify patients who may have been misdiagnosed with other conditions. This AI, developed under Project Zebra, significantly reduced the diagnostic time by an average of 1 to 2 years, thereby improving patient outcomes [18]. The DeepMind team developed an AI model named AlphaMissense to predict whether DNA mutations are benign or pathogenic. This AI helps to identify potentially harmful mutations in genes linked to rare diseases, aiding in faster and more accurate diagnoses [19]. Upon examining these and similar examples, it can be seen that AI can become a successful assistant to doctors in medical matters, with various parameters considered. This study attempts to demonstrate the stance of AI in terms of diagnosis and detection through the musculoskeletal system.

As is well known, the primary function of the musculoskeletal system is to protect the body's organs from external mechanical effects, provide support, and enable mobility [20]. An adult human body consists of 206 bones. The structural integrity of bones can be compromised by external forces or traumas, resulting in simple fractures, multiple fractures, or even open fractures where the bone protrudes outside the body [21]. In the field of orthopedics and traumatology, proximal femur fractures are among the leading causes of emergency interventions and hospitalizations, particularly in individuals over the age of 65. The mortality rate after a hip fracture remains significant, being identified as 11–23 % within 6 months and 22–29 % one year after the injury [22, 23]. When examining all proximal femur fractures, more than 90 % of these fractures are found to be femoral neck fractures, with the remaining portion consisting of subtrochanteric fractures [24, 25]. Approximately 98 % of femoral proximal fractures occur in patients over the age of 50 [26].

For this study, proximal femoral fractures, which are frequently observed in the elderly population and have the highest disability and mortality rates, were chosen to compare the behavioral approaches of doctors and artificial intelligence. As a sample for the study, one of the largest hospitals in Istanbul with an orthopedic and traumatology department representing the general population of Turkey was selected. An investigation was conducted on 500 patients admitted to this hospital due to hip trauma. More than 1 500 images of the femur upper region were obtained from these 500 patients. Initially, the algorithms capable of detecting this specific trauma, YOLO Darknet v4, Faster R-CNN – Inception v2, and SSD – MobileNet v2,

were trained. Two different physician groups were formed to compare the behaviors of these retrained AI models. The first physician group consisted of 10 general practitioners, and the second group comprised 12 expert doctors and senior residents from the orthopedic and traumatology department.

To enable an independent evaluation and to prevent the AI models from memorizing biases or outcomes, a total of 50 different evaluation datasets were created, independent of previously collected data; 50 datasets contained femoral proximal fractures, while 55 datasets contained no fractures. The outcomes produced by each segregated physician group and each segregated AI group for each evaluation dataset were analyzed using one-way ANOVA and Tukey pairwise tests to determine whether there were similarities in their approaches in this context.

2 LITERATURE RESEARCH

Mankind, on its journey to analyze the root causes of its own behaviors, has now begun to discuss the similarities and differences between artificial intelligence (AI) behaviors and human intelligence. In 2021, Koterling and his colleagues conducted a study analyzing the fundamental differences and similarities between humans and AI. While highlighting humans' intuitive and creative thinking abilities, the study pointed out that AI excels in big data and complex computations. It also emphasized that AI plays a critical role in complementing human capabilities [27]. In 2021, Triberti and colleagues emphasized that the successful implementation of artificial intelligence depends not only on its technical effectiveness but also on how people perceive, understand, accept, and use AI-based systems. They identified users' attitudes and behaviors, the design and validation of human–AI interfaces, explainable AI, human factors, and individual characteristics that may facilitate or hinder interaction as important areas for further research. The editorial therefore highlighted the need for a human-centered approach in which AI technologies are developed and evaluated by considering users' needs, motivations, limitations, and real-world contexts [28].

In 2024, a study by Thirunavukarasu and colleagues at the University of Cambridge examined the accuracy of large language models in clinical knowledge and assessments of eye diseases. The study observed that AI models approached the level of human experts, especially in clinical knowledge and reasoning when compared with ophthalmologists. In an 87-question exam, GPT-4 achieved a 69 % accuracy rate, GPT-3.5 achieved 48 %, LLaMA achieved 32 %, and PaLM 2 achieved 56 %, while expert ophthalmologists achieved 76 %, ophthalmology residents 59 %, and general practitioners 43 % [29]. In 2023, Ippolito and colleagues conducted a study using AI to diagnose lung pneumonia in the emergency department. Expert radiologists achieved 96 % sensitivity and 79.8 % specificity in detecting COVID-19, while the AI system achieved 94.7 % sensitivity and 80.2 % specificity. In detecting pneumonia, radiologists achieved a success rate of 97.9 % sensitivity and 88 % specificity, while the AI system achieved 97.5 % sensitivity and 83.9 % specificity [30].

In 2019, Rahwan and colleagues, in their study titled "Machine Behavior", examined how AI acts and makes decisions independently. The study emphasized that AI systems could exhibit unpredictable and complex behaviors, which necessitates new methods for researchers to understand these behaviors. The study also addressed the social impacts and ethical issues of AI systems. It underscored the need for more research to understand the impacts of AI on humans and society. Additionally, the study highlighted the need to examine AI's learning processes and solve the issues of explainability [31]. In 2020, Pedersen and Johansen introduced behavioural artificial intelligence as a research agenda for investigating how intelligent systems form judgments and reach decisions. Rather than assuming that AI systems reason in the same way as humans or arguing that they should acquire human-like capabilities, the authors called for systematic empirical studies of artificial inference using perspectives and methods from the cognitive and behavioural sciences, natural sciences, and philosophy. They argued that such research could improve understanding of the often complex and opaque processes underlying AI-generated judgments and decisions. This knowledge would also support meaningful human oversight and contribute to greater accountability in the use of intelligent systems [32]. Dietvorst and colleagues, in their 2015 study, explored how individuals lose trust in machines/algorithms when they make mistakes, leading them to avoid using these algorithms. The study showed that there is a common misconception among people that algorithms must be perfect. It concluded that people abandon using algorithms even if they perform better than humans when they make errors. The study emphasized the importance of education and awareness to enhance the acceptance and trust in algorithms [33].

In 2014, Cummings emphasized the superiority of human-machine collaboration over individual performance. It demonstrated that cooperation between humans and machines yielded better results than human or machine performance alone. It was found that human decisions aided by machines were faster and more accurate. The study also highlighted that automation could reduce workload on humans and improve decision-making processes. However, it noted that trust and proper design are critical for the success of human-machine collaboration [34]. De Visser and Parasuraman's 2011 study examined the performance and trust of human-robot teams, investigating how automation affects performance even when it is not perfect. The study concluded that performance and trust could be enhanced even when automation is not perfect. It observed that when people understand the limitations of automation and develop strategies to cope with these limitations, they improve their performance. Moreover, it was found that automation reduces workload and supports decision-making processes [35]. Silver and his colleagues demonstrated in 2016 how artificial intelligence could skillfully play the game of Go by using deep neural networks and tree search methods. AlphaGo's successes against world champion human players have proven that artificial intelligence can perform excellently in complex strategic games. As a result, they revealed that AI could compete with human intuition and creative strategies. This success highlights the power and potential of deep learning and search algorithms [36].

In 2013, McDuff and colleagues worked on artificial intelligence in the field of emotional intelligence. The study introduced the Affectiva-MIT facial expression dataset (AM-FED) to collect and analyze natural and spontaneous facial expressions in everyday settings. The results showed that artificial intelligence could make significant progress in the field of emotional intelligence. The study aimed for AI to accurately detect people's emotional states and use this information to provide more effective and empathetic interactions. This study was an important step toward developing AI's social and emotional intelligence capabilities [37]. In 2011, Hancock and colleagues examined factors that affect trust in human-robot interactions. The study found that the characteristics of robots, such as reliability, ability, and predictability, increased people's trust in robots. It also found that robots behaving similarly to humans and effectively communicating with humans also increased trust. These findings provide important clues for reliable and effective human-robot collaboration [38].

In 1996, Nass and colleagues posed the question of whether computers can be teammates. The study examined whether computers could be accepted as human teammates. The results showed that people perceived computers as social entities in their interactions with them and treated them similarly to how they treat humans. It was stated that computers could be accepted as teammates by humans if they are reliable, predictable, and cooperative. These findings reveal the social dimension of human-computer interaction and the potential for artificial intelligence to collaborate more effectively with humans [39]. In 2017, Bryson and Winfield addressed ethical design standards for artificial intelligence and autonomous systems. The study emphasized the importance of ethical design of AI systems and argued that these systems should be designed in a trustworthy, fair, and responsible manner. It was stated that transparency, accountability, and ethical principles should be applied in the decision-making processes of AI. This is critical to increasing the social acceptance of AI and reducing its negative impacts [40].

In 2019, Yeomans and colleagues studied the impact of AI-based recommendation systems on human judgments. The study focused on how AI recommendation systems shape users' decision-making processes and how these recommendations are perceived. The study found that AI recommendations are generally valued by users, but the explainability and transparency of these recommendations are critical to increasing user trust. Users tend to accept recommendations when they understand the reasoning behind them [41]. Luckin and colleagues, in their 2016 study, examined the potential of artificial intelligence in education. The study demonstrated that AI offers personalized learning experiences to users and responds better. It was found that AI-supported educational tools help teachers develop more effective teaching methods by providing support. This study showed that AI has the power to transform education and play a significant role in increasing student success [42]. Colton and Wiggins, in their 2012 study, explored the role of artificial intelligence in creative processes. The study examined how AI could innovate in creative fields such as music, art, and writing, and how it could be compared with human creativity. The study concluded that AI could inspire people in creative processes

and produce new creative works. It was stated that AI has significant potential in creative industries and can play a complementary role [43]. In 2019, Floridi and Cowls aimed to regulate the impact of artificial intelligence on society through their study, focusing on five principles: beneficence, non-maleficence, autonomy, justice, and transparency. They emphasized that AI should be designed ethically and responsibly, and that these principles should be the foundation for the development of AI systems. The importance of adhering to these ethical principles was emphasized to ensure that AI systems provide benefits to society [44].

Esteva and colleagues conducted a comprehensive study in 2017 on skin cancer classification using deep neural networks. Their models showed results equivalent to dermatologists' performance with 95 % sensitivity and 94 % specificity [45]. In 2017, Lakhani and Sundaram focused on the use of convolutional neural networks (CNN) to classify lung tuberculosis in chest X-rays. Their study revealed that the AI system performed similarly to human radiologists with 97.3 % accuracy, 98.1 % sensitivity, and 96.2 % specificity [46]. In 2018, Rajpurkar and colleagues developed the CheXNeXt algorithm for chest X-ray diagnosis and compared the algorithm's performance with practising radiologists. The developed AI algorithm achieved accuracy levels comparable to radiologists for certain diseases, with 96.8 % overall accuracy, 93.5 % sensitivity, and 95.2 % specificity [47]. In 2016, Gulshan and colleagues created a deep learning algorithm to detect diabetic retinopathy from retinal fundus photographs. The algorithm reached 97.5 % sensitivity and 93.4 % specificity, performing at the same level as expert ophthalmologists [48]. Liu and colleagues' 2019 study stated that the AI system detected intracranial hemorrhages with 87.2 % accuracy, while human radiologists had an accuracy rate of 81.3 % [49]. In their 2023 study, Zeren and colleagues compared the performance of medical professionals in detecting proximal femur fractures in X-ray images with the YOLO algorithm. The YOLO algorithm was successful with an accuracy rate of 90.3 %, while orthopedic specialist doctors achieved a success rate of 91.4 %, and general practitioners had an accuracy rate of 81.3 % [50]. Liu and colleagues' 2019 study reviewed studies published between 2000 and 2019 that compared advanced AI technologies with doctors. The research commonly used convolutional neural networks as the AI technology. It compared the diagnostic accuracy, false positive rate, sensitivity, and specificity between AI and doctors. The combined results revealed that AI diagnostic performance was similar to that of experienced doctors and better than that of inexperienced doctors [51]. In 2020, Baker and colleagues conducted a prospective validation study comparing the performance of the Babylon Triage and Diagnostic System with that of human doctors. The researchers used a semi-naturalistic role-playing paradigm based on simulated clinical cases to evaluate history-taking, differential diagnostic performance, and the safety and appropriateness of triage recommendations. The AI system generated differential diagnoses with precision and recall broadly comparable to those of the doctors and exceeded human performance in some assessments. It also produced safer triage recommendations on average than the participating doctors, although the appropriateness of its recommendations was marginally lower. These findings suggest that the evaluated AI system could pro-

vide diagnostic and triage information at a level comparable to human doctors under simulated conditions; however, further validation in real-world clinical settings is required [52]. In 2019, Bhuva and colleagues compared a fully automated convolutional neural network with an expert and a trained junior clinician for the analysis of cardiovascular magnetic resonance images. Using scan–rescan data from 110 patients representing five disease categories and five institutions, the researchers evaluated the precision of measurements relating to left ventricular volume, mass, and ejection fraction. The automated method demonstrated scan–rescan precision similar to that of the human observers while completing the analysis approximately 186 times faster, requiring about four seconds rather than 13 minutes. These findings indicate that machine-learning-based cardiac MRI analysis may substantially improve efficiency while maintaining measurement precision comparable to human analysis. However, the study assessed the quantitative analysis of cardiac imaging biomarkers rather than the direct diagnostic accuracy of AI in identifying heart disease [53]. In 2019, Haenssle and colleagues conducted a study comparing the accuracy of a deep learning convolutional neural network (CNN) with 58 dermatologists in diagnosing dermoscopic melanoma. The AI system performed better than dermatologists in terms of specificity [54].

3 METHOD

In this study, two physician groups consisting of orthopedic specialists and general practitioners were examined alongside three deep learning algorithm groups: YOLO Darknet V4, SSD MobileNet v2, and Faster R-CNN Inception v2 algorithms. These deep learning algorithms are the best for object detection, especially on complicated and distorted images. And their architecture is different. An independent test dataset, produced separately from the training and validation datasets, was created to assess whether there was a statistically significant difference between the results generated by these groups in terms of time taken to produce results, accuracy in detecting fractures in broken bones, and accuracy in detecting non-broken bones. The stopwatch method was used to calculate result production times. During the stopwatch method, the procedure also recorded potential human errors. This method involves measuring the time required for each step or activity with a stopwatch to determine the durations necessary to complete these steps. It is widely used to increase the efficiency of work processes, plan labor, and establish job standards [55, 56, 57, 58].

The one-way ANOVA test, also known as a single-factor analysis of variance, is used to check whether there is a statistically significant difference between the means of the results produced by independent groups. ANOVA stands for “Analysis of Variance”. Unlike a two-sample t-test, which compares groups in pairs, a one-way ANOVA test controls the relationships of multiple groups at once. Provides a cumulative comparison of the arithmetic means of three or more data groups. The one-way ANOVA test is a parametric test that requires the data to be normally

distributed and the sample size to be sufficient for use. To quickly obtain results when performing a one-way variance analysis, MiniTab or SPSS programs are often used. Regardless of the program used, if the p-value in a one-way ANOVA test is less than 0.05, the differences between the means are considered significant. H0: The means of all series are equal; H1: At least one series has a mean that is significantly different from the others; if the p-value is less than 0.05, H0 is rejected, and H1 is accepted. Studies on one-way variance analysis show that it is a test not sensitive to normality, and analyses can be conducted if the sample size is sufficient.ANOVA test is known to be robust to moderate violations of normality, particularly when sample sizes are sufficiently large. Given the sample size in this study, minor deviations from normality were unlikely to impact the validity of the findings significantly [59].

Total Sum of Squares for Variance (SST):

$$\text{SST} = \sum_{i=1}^{N} (Y_i - \overline{Y})^2, \tag{1}$$

where Y_i is the value of each observation, $\overline{Y}$ is the mean of all observations, N is the total number of observations.

Between-Group Sum of Squares (SSB):

$$\text{SSB} = \sum_{j=1}^{k} n_j (\overline{Y_j} - \overline{Y})^2, \tag{2}$$

where $\overline{Y_j}$ is the mean of group j, n_j is the number of observations in group j, k is the total number of groups.

Within-Group Sum of Squares (SSW):

$$\text{SSW} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\overline{Y}_{ij} - \overline{Y_j})^2, \tag{3}$$

where $\overline{Y}_{ij}$ is the i^{th} observation value in group j, $\overline{Y}_j$ is the mean of group j.

- Degrees of Freedom (df): The degrees of freedom are calculated for each source of variance.
- Between-group degrees of freedom: $df_{between} = k - 1$.
- Within-group degrees of freedom: $df_{within} = N - k$.
- Total degrees of freedom: $df_{total} = N - 1$.
- Mean Squares (MS):
  - Between-group mean squares: $MS_{between} = \frac{SSB}{df_{between}}$.
  - Within-group mean squares: $MS_{within} = \frac{SSW}{df_{within}}$.

- F-Statistic: $F = \frac{MS_{between}}{MS_{within}}$.

The p-value is used to determine where the F-statistic falls within the distribution. The p-value represents the probability of obtaining the observed F-statistic by chance. The p-value can be found from the F-distribution table based on the calculated F-statistic and degrees of freedom [60, 61].

While one-way variance analysis shows whether there are differences between the groups, it does not provide information about which groups are different. Therefore, multiple comparison tests are continued. Since variances are similar and sample sizes are equal, Tukey's comparison tests are continued. Tukey's test, developed by J. W. Tukey in 1949, is a multiple comparison test that enables pairwise comparisons of group means using a common error approach for unrelated groups. The important reasons for choosing the Tukey test can be classified as follows:

True Significant Differences: The Tukey test is used to determine the true significant differences among multiple comparisons. This means it is a robust method for identifying significant differences between groups.

Controls Excessive Type I Error: When conducting multiple comparisons, performing multiple tests using the significance level usually set for a single hypothesis test (typically alpha = 0.05) increases the risk of excessive Type I error. The Tukey test adjusts the significance level to keep this risk under control.

Eases the Homogeneity Assumption: The Tukey test can be used in cases where groups are not homogeneous, as a result of variance analyses like ANOVA. Therefore, reliable results can be obtained even when variances are not homogeneous.

Manages the Number and Size of Groups Well: The Tukey test is flexible concerning the number of groups and sample sizes. It can be used with different group numbers and sample sizes.

Simple and Structured Results: The results of the Tukey test are usually simple and clear. It clearly shows which groups are significantly different from each other.

Suitable for Various Data Distributions: The Tukey test is not dependent on the normal distribution assumption and is quite tolerant of different data distributions [59, 62, 63].

In the Tukey HSD Test:
Calculation of Honestly Significant Difference (HSD):

$$HSD = q\sqrt{\frac{MS_{within}}{n}}, \tag{4}$$

where q is the critical value obtained from Tukey's distribution (studentized range distribution), MS_{within} is the within-group mean squares obtained from the ANOVA results, n is the sample size for each group.

Pairwise Comparison of Mean Differences:
The mean difference between two groups:

$$\Delta\overline{Y}_{ij} = |\overline{Y}_i - \overline{Y}_j|, \tag{5}$$

where $\overline{Y}_i$ and $\overline{Y}_j$ are the means of groups i and j, respectively.

In the Tukey HSD test results, for all pairwise comparisons between groups, the mean differences and their confidence intervals are calculated. If the mean difference of a pair is greater than the HSD value, it is concluded that the difference between these two groups is statistically significant [60, 61].

Before subjecting the given result series to one-way ANOVA (variance) significance tests, it is necessary to check for normal distribution. To check the conformity to normal distribution, it is important to look at the skewness and kurtosis values of each series separately.

Skewness represents how much the data distribution is skewed to the right or left. Positive Skewness: The distribution is skewed to the right (long tail on the right). Negative Skewness: The distribution is skewed to the left (long tail on the left). Skewness = 0: The distribution is symmetrical.

Kurtosis is a measure that indicates how peaked or flat the data distribution is. Kurtosis > 3: The distribution has a higher peak and longer tails than normal (positive kurtosis). Kurtosis < 3: The distribution has a flatter peak and shorter tails than normal (negative kurtosis). Kurtosis = 3: The distribution is normal (standard kurtosis).

Tabachnick and colleagues' 2013 study indicated that if the skewness and kurtosis values of the series range between -1.5 and +1.5, it can be considered normally distributed, and George and Mallery's 2010 study suggested that if the skewness and kurtosis values range between −2.0 and +2.0, it can be considered normally distributed [64, 65].

$$\text{Skewness} = \frac{n\sum(x_i - \overline{x})^3}{(n-1)(n-2)s^3}, \tag{6}$$

$$\text{Kurtosis} = \frac{n(n+1)\sum(x_i - \overline{x})^4 - 3(n-1)^2 s^4}{(n-1)(n-2)(n-3)s^4}. \tag{7}$$

In the above formulas, n is Number of observations, x_i is i^{th} observation value, $\overline{x}$ is Mean, s is Standard deviation.

4 APPLICATION

In this study, deep learning and computer vision algorithms considered among the most powerful in their fields were selected to compare against well-trained specialist doctors and general practitioners. The primary objective was to complete transfer learning and create training and validation datasets to develop AI models specialized only in this field. For this purpose, X-ray images of over 500 patients

who applied to the orthopedic or emergency clinic at a training and research hospital in Istanbul – Türkiye after experiencing trauma were examined. The patients' origin was Turkish. X-ray images without distortions or problems were separated from the others. A dataset containing X-ray images of the femur upper region from 410 patients was selected to create training and validation datasets. 72 % of patients were women and 23 % of patients were men. 72.4 % of the patients were older than 70 years old. In addition to these 410 patients, 50 femur upper region X-rays with fractures and 50 without fractures, making a total of 55 independent test datasets, were separated from all other data. The dataset consisted of patients who visited the hospital in the last three years, aged between 22 and 105. The Collection of validation datasets includes the same number of intertrochanteric fractures and collum femoris fractures, which are type of proximal femur fractures. During the collecting process validation dataset complex and non-complex cases balanced by research team which includes specialist doctors. All data sets' X-ray images were obtained from a single imaging protocol and system. also for all dataset annotations and labels assigned by multiple radiologists and multiple specialist doctors. During the labeling process, there was no disagreement. If the research team had encountered such a case, the research team would have applied the second opinion of the professor doctors who were consultants.

The training and evaluation dataset consists of femur upper region X-ray images obtained from 410 patients (Figure 1). Using data augmentation techniques, including rotation and scaling, the dataset was doubled to a total of 820 images. To prevent overfitting in these deep learning and computer vision algorithms, data augmentation techniques used and also diverse data used, a larger dataset with diverse conditions improved the model's generalization. During the training validation loss tracked and the training stopped according to that.

Figure 1. Examples of the created dataset

In the study, transfer learning was completed to develop the YOLO – You Look Only Once – Darknet V4 algorithm, SSD – Single Shot Multibox Detector – MobileNet v2 algorithm, and the most powerful R-CNN algorithm, Faster R-CNN –

Inception v2, for each dataset by re-marking the coordinates of the fractured femur upper region in text format (Figure 2).

Figure 2. Marking the fractured areas in text format

For all three algorithms – YOLO Darknet V4, SSD MobileNet v2, and Faster R-CNN Inception v2 – the Python programming language was chosen due to its extensive use in both library selection and software language. During the selection of the user interface (Idle), the Spider and PyCharm interfaces, which provide significant convenience and shortcuts to the user, were chosen.

In this study, particularly for the YOLO Darknet V4 algorithm, the cloud technology of software as a service (SaaS) was preferred due to the anticipation that the training process would be quite lengthy on any workstation CPU or GPU. During the selection of cloud technology, the aim was to retrain the model with one of the world's most powerful GPUs, the Tesla K80 24 GB GDDR5 GPU, belonging to Google Colaboratory. For the retraining of the SSD MobileNet v2 and Faster R-CNN Inception v2 algorithms, a workstation considered to be powerful, Intel(R) Core(TM) i5-8300H CPU @ 2.30 GHz – 16 GB RAM – GTX1050 GPU, was selected.

After retraining the YOLO Darknet V4, SSD MobileNet v2, and Faster R-CNN Inception v2 algorithms, a powerful workstation (Intel(R) Core(TM) i5-8300H CPU @ 2.30 GHz – 16 GB RAM – GTX1050 GPU) was again chosen to produce results on the independent test dataset.

One of the critical stages of the research was the determination and grouping of general practitioners and orthopedic specialists. Orthopedic specialist doctors were selected from one of the most significant training and research hospitals in the orthopedics and traumatology service in Istanbul, the city with the largest population in Turkey. A total of 12 specialist doctors and senior assistants from the orthopedics and traumatology service were focused on. They were presented with a mixed set of 100 femur upper region images, consisting of 50 femur upper region images with fractures and 50 without fractures, created as a completely independent test dataset of 55 X-ray images. Whenever an X-ray image was shown to them, they were expected to indicate whether there was a right femur upper region fracture, a left femur upper region fracture, both a right and left femur upper region fracture, or no fracture at all. The times from the moment the doctors first saw the image to the moment they provided their response by marking the fracture on the image were measured using the stopwatch method and recorded as result production

times. The results were recorded in this way, and calculations and measurements were made according to the evaluation criteria table.

While selecting general practitioners, doctors who had spent more than one month in different branches of an educational and research hospital in Istanbul, the most populous city in Türkiye, emergency doctors, and family doctors from different locations were included. The study was conducted with a total of 10 general practitioners. They were shown a mixed set of 100 femur upper region images, consisting of 50 femur upper region images with fractures and 50 without fractures, created as a completely independent test dataset of 55 X-ray images. As with the specialist doctors, whenever an X-ray image was shown, they were expected to indicate whether there was a right femur upper region fracture, a left femur upper region fracture, both a right and left femur upper region fracture, or no fracture at all. The times from the moment the doctors first saw the image to the moment they provided their response by marking the fracture on the image were again measured using the stopwatch method and recorded as result production times.

With this study, datasets containing the results for a total of 55 different X-ray images, 50 with fractures in the femur upper region and 50 healthy femur upper regions, were created. These datasets included the result production times for the femur upper region, the accuracy rates for the fractured region, and the accuracy rates for the non-fractured region. The first series, the femur upper region result production times series, was expected to contain a total of 275 results, with 55 for each study group. The accuracy rates series for the fractured femur upper region was expected to contain a total of 250 results, with 50 for each study group. Similarly, the accuracy rates series for the non-fractured femur upper region was expected to contain a total of 250 results, with 50 for each study group.

5 RESULTS

The results of both, the three different artificial intelligence models obtained through transfer learning and the two different physician groups, were evaluated using Minitab and SPSS programs. Afterwards, the outputs produced were assessed to determine if there was a relationship between them using one-way ANOVA tests and Tukey pairwise tests.

The femur upper region result production time series, which consists of a total of 275 data points, the accuracy rate series for the fractured femur upper region, consisting of 250 data points, and the accuracy rate series for the non-fractured femur upper region, consisting of 250 data points, were calculated for each study group using the SPSS program. The skewness and kurtosis values obtained are shown in Figures 3, 4 and 5.

The results from the analysis of three different data groups showed that the skewness and kurtosis values for femur upper region result production times, fractured femur upper region accuracy, and non-fractured femur upper region accuracy ranged between -2 and +2, indicating a normal distribution. Consequently, the se-

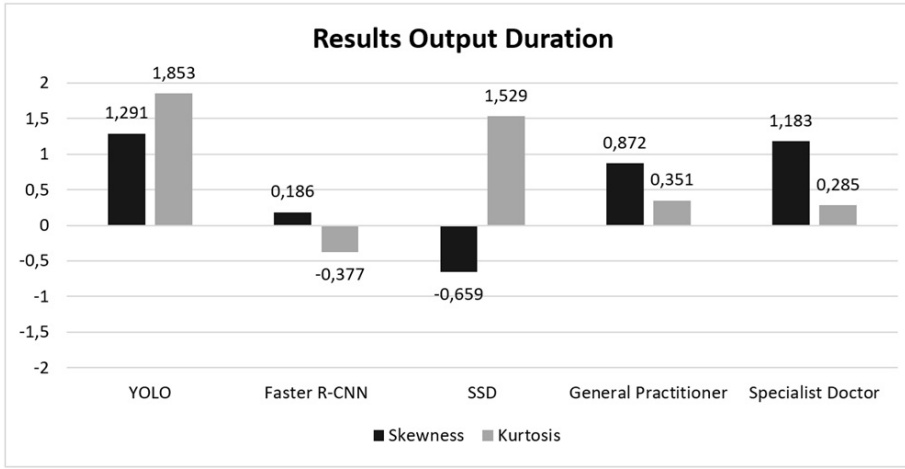


Figure 3. Skewness and kurtosis analysis results for femur upper region result production times

ries were individually subjected to one-way ANOVA tests, followed by Tukey tests to identify the pairwise relationships.

Regarding femur upper region result production times, orthopedic specialists, general practitioners, YOLO algorithm, Faster R-CNN algorithm, and SSD algorithm were subjected to one-way ANOVA tests, and when analyzed using the MiniTab program, the results shown in Figure 6 were obtained.

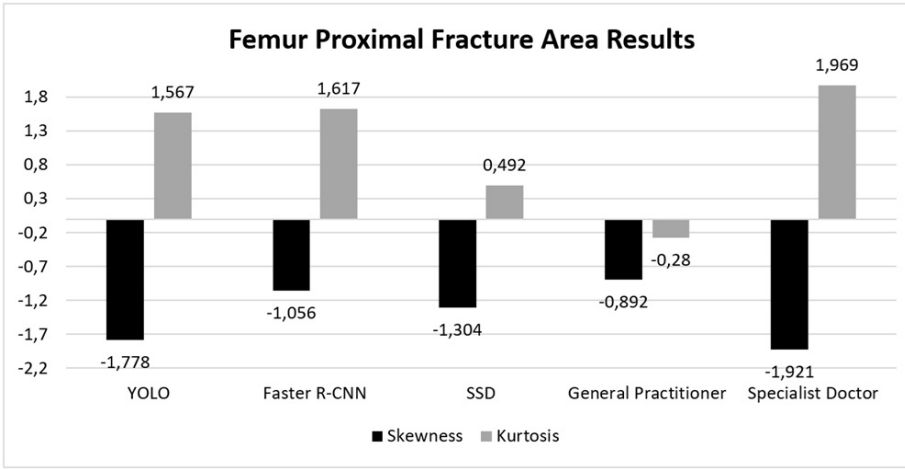


Figure 4. Skewness and kurtosis analysis of accuracy in the fractured femur upper region

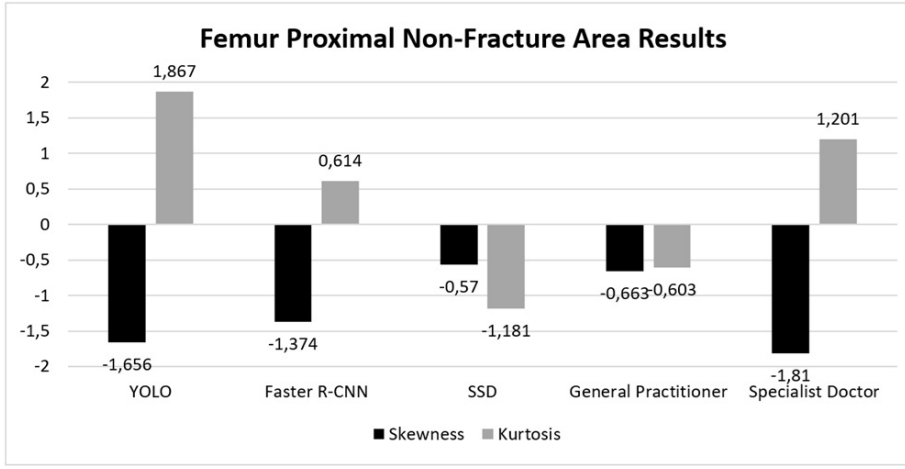


Figure 5. Skewness and kurtosis analysis of accuracy in the non-fractured femur upper region

WORKSHEET 1

One-way ANOVA: Specialist Doctor; General Practitioner; YOLO; Faster-RCNN; SSD

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0,05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Factor	5	Specialist Doctor; General Practitioner; YOLO; Faster-RCNN; SSD

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	4	2375,3	593,817	228,26	0,000
Error	270	702,4	2,602		
Total	274	3077,7			

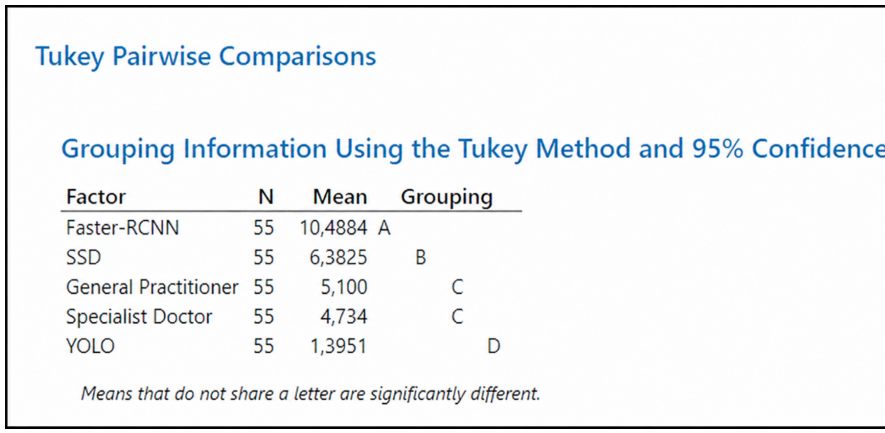
Figure 6. One-way ANOVA analysis of femur upper region result production times (DF: Degrees of Freedom, Adj SS: Adjusted Sum of Squares, Adj MS: Adjusted Mean Square)

Since the p-value is less than 0.05, there is a significant difference between the mean result production times. Based on this outcome, H0 is rejected, and H1 is accepted.

H0: All the mean result production times are equal.

H1: At least one mean is different from the others.

When Tukey pairwise comparison tests, which are among the pairwise comparison tests, were continued using the MiniTab program, the results shown in Figure 7 were obtained.



Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping		
Faster-RCNN	55	10,4884	A		
SSD	55	6,3825		B	
General Practitioner	55	5,100			C
Specialist Doctor	55	4,734			C
YOLO	55	1,3951			D

Means that do not share a letter are significantly different.

Figure 7. Tukey pairwise comparison analysis of femur upper region results production times

According to Tukey pairwise analysis, the difference between means with the same letter group is insignificant, while the difference between means with different letter groups is significant. Accordingly, the average result production time for the Faster R-CNN Algorithm is significantly higher than for the SSD Algorithm, general practitioners, orthopedic specialists, and YOLO Algorithm. In contrast, the average result production time for general practitioners is not significantly different from that of orthopedic specialists. The average result production time for orthopedic specialists is significantly higher than that of both the SSD algorithm and the Faster R-CNN algorithm.

Furthermore, the YOLO algorithm's average result production time is significantly lower. Accordingly, the result production time of the expert physicians is similar to that of other artificial intelligences trained through transfer learning. The thinking and analysis times in result production exhibit similar behavioral characteristics.

Regarding the accuracy rate for the fractured femur upper region, orthopedic specialists, general practitioners, YOLO algorithm, Faster R-CNN algorithm, and

SSD algorithm were subjected to one-way ANOVA tests, and when analyzed using the MiniTab program, the following results were obtained:

WORKSHEET 1

One-way ANOVA

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0,05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Factor	5	YOLO accuracy in the fractured femur upper region; Faster-RCNN accuracy in the fractured femur upper region; SSD accuracy in the fractured femur upper region; General Practitioner accuracy in the fractured femur upper region; Specialist Doctor accuracy in the fractured femur upper region

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	4	1,157	0,28917	4,84	0,001
Error	245	14,642	0,05976		
Total	249	15,799			

Figure 8. One-way ANOVA analysis of accuracy in the fractured femur upper region (DF: Degrees of Freedom, Adj SS: Adjusted Sum of Squares, Adj MS: Adjusted Mean Square)

Since the p-value is less than 0.05, there is a significant difference between the mean accuracy rates for fractured regions. Accordingly, H0 is rejected, and H1 is accepted.

H0: All the mean success rates are equal.

H1: At least one mean is different from the others.

Following the one-way variance tests, Tukey pairwise comparison tests were continued, and the results shown in Figure 8 were obtained using the MiniTab program.

According to Tukey pairwise analysis, the difference between means with the same letter group is insignificant, while the difference between means with different letter groups is significant. Accordingly, the success rate of orthopedic specialists in fractured region accuracy is significantly higher than that of the SSD algorithm and general practitioners, but the difference is insignificant compared to the Faster R-CNN and YOLO Algorithms. The success rate of orthopedic specialists in detecting fractures in fractured regions is significantly higher than that of the SSD algorithm and is said to show similar behavioral characteristics for the same test dataset. The

Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping		
Specialist Doctor accuracy in the fractured femur upper region	50	0,9167	A		
Faster-RCNN accuracy in the fractured femur upper region	50	0,9064	A	B	
YOLO accuracy in the fractured femur upper region	50	0,8767	A	B	C
SSD accuracy in the fractured femur upper region	50	0,7832		B	C
General Practitioner accuracy in the fractured femur upper region	50	0,7480			C

Means that do not share a letter are significantly different.

Figure 9. Tukey pairwise comparison analysis of accuracy in the fractured femur upper region

success rate of general practitioners in detecting fractures in fractured regions is significantly lower than that of the Faster R-CNN algorithm. It can be said to show similar behavioral characteristics in similar test images. However, the YOLO algorithm is seen to exhibit a behavior distinct from all groups, indicating a unique ability to produce correct results for the fractured region.

Regarding the accuracy rate for the non-fractured femur upper region, orthopedic specialists, general practitioners, YOLO algorithm, Faster R-CNN algorithm, and SSD algorithm were subjected to one-way ANOVA tests, and when analyzed using the MiniTab program, the results shown in Figure 10 were obtained.

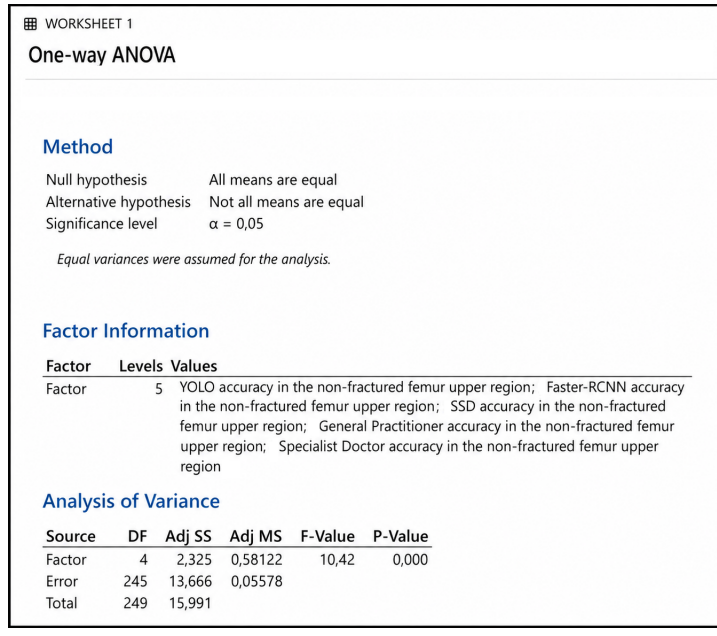
Since the p-value is less than 0.05, there is a significant difference between the mean accuracy rates for non-fractured regions. Accordingly, H0 is rejected, and H1 is accepted.

H0: All the mean success rates are equal.

H1: At least one mean is different from the others.

Following the one-way variance tests, Tukey pairwise comparison tests were continued, and the results shown in Figure 11 were obtained using the MiniTab program.

According to Tukey pairwise analysis, the difference between means with the same letter group is insignificant, while the difference between means with different letter groups is significant. Accordingly, the success rate of the YOLO Algorithm in non-fractured region accuracy is significantly higher than that of the Faster R-CNN and SSD Algorithms, but the difference is insignificant compared to orthopedic specialists and general practitioners. The success of orthopedic specialists in non-fractured regions is significantly lower than that of both the Faster R-CNN and SSD algorithms. Therefore, it can be said that the fracture detection success of



WORKSHEET 1

One-way ANOVA

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0,05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Factor	5	YOLO accuracy in the non-fractured femur upper region; Faster-RCNN accuracy in the non-fractured femur upper region; SSD accuracy in the non-fractured femur upper region; General Practitioner accuracy in the non-fractured femur upper region; Specialist Doctor accuracy in the non-fractured femur upper region

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	4	2,325	0,58122	10,42	0,000
Error	245	13,666	0,05578		
Total	249	15,991			

Figure 10. One-way ANOVA analysis of accuracy in the non-fractured femur upper region (DF: Degrees of Freedom, Adj SS: Adjusted Sum of Squares, Adj MS: Adjusted Mean Square)

orthopedic specialists in the test dataset for non-fractured regions shows similar behavioral characteristics to those of the Faster R-CNN and SSD algorithms. However, the success of general practitioners in non-fractured regions is significantly higher than that of both the Faster R-CNN and SSD algorithms. In terms of accuracy in non-fractured regions, the YOLO algorithm distinguishes itself from both physician groups.

6 SUGGESTIONS AND DISCUSSIONS

This study focused on proximal femoral fractures, one of the orthopedic fractures with a high rate of disability and mortality in the elderly population worldwide, to reveal the similarities in behaviors and output production between AI models retrained to specialize in this field and experts in the field. Along with this study, the similarity of AI in producing outputs and the scope of the efficiency and added value that joint work will provide have been demonstrated.

It has been understood that AI algorithms have great potential to support and accelerate the diagnostic processes of specialists. More reliable results can be obtained if AI systems are used not entirely independently but in conjunction with the experience and knowledge of specialists. Especially in complex cases and rare

Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping		
YOLO accuracy in the non-fractured femur upper region	50	0,9298	A		
Specialist Doctor accuracy in the non-fractured femur upper region	50	0,9117	A		
General Practitioner accuracy in the non-fractured femur upper region	50	0,8780	A	B	
Faster-RCNN accuracy in the non-fractured femur upper region	50	0,7794		B	C
SSD accuracy in the non-fractured femur upper region	50	0,6716			C

Means that do not share a letter are significantly different.

Figure 11. Tukey pairwise comparison analysis of accuracy in the non-fractured femur upper region

conditions, when the clinical knowledge and experience of specialists are combined with the analyses provided by AI, diagnostic accuracy will increase. Such a hybrid approach will increase patient safety and enhance the quality of healthcare services. Furthermore, the interaction between specialists and AI offers the opportunity to optimize the performance of both parties. It is essential to provide the necessary training for healthcare professionals and specialists to use AI technologies effectively. Comprehensive training programs should be developed on how AI systems work, how data is processed, and how results are interpreted. In addition to training programs, continuous professional development opportunities and hands-on training should be provided to support healthcare workers' adaptation to AI systems. Moreover, ethical and legal issues should also be addressed during the training process. Such as AI models may inherit biases from training data, leading to disparities in healthcare outcomes; there could be risk of discrimination against certain demographic groups; AI decisions may lack transparency, making it difficult for patients to provide informed consent; Patients should have the right to understand and question AI-driven diagnoses; Compliance with data protection laws is crucial to prevent misuse; AI-driven medical devices and software must meet regulations like the FDA (USA), MDR (EU), and MHRA (UK) before being deployed; Determining liability in case of AI errors is complex (e.g., doctor, hospital, AI developer); Laws may need updates to define responsibility clearly.

Studies in the field of AI and medical diagnostics should be continued, and the accuracy of the algorithms should be increased with more extensive datasets. More studies are needed to evaluate the performance of AI systems on different medical conditions and diseases. This will not only contribute to improving existing AI systems but also to the development of new algorithms. Additionally, studies evaluating the performance of AI systems on different populations and demographic groups are also needed. Such studies will enhance the generalizability and universal

applicability of AI systems. Healthcare professionals should be enabled to use AI technologies consciously and responsibly. The use of AI systems in clinical practice requires ongoing feedback and improvement processes. Feedback from healthcare professionals and specialists on the results obtained from AI systems is critical to further improving these systems. Challenges and successes encountered in clinical practice will allow AI systems to adapt to real-world conditions. Therefore, a robust collaboration and communication network should be established between healthcare institutions and AI developers. To use AI technologies effectively in the healthcare sector, multidisciplinary approaches should be adopted. Collaboration among computer scientists, data scientists, biomedical engineers, and clinicians will enable the development of AI systems more effectively and efficiently. Moreover, multidisciplinary research projects and joint studies will bring together different perspectives, allowing innovative solutions to be developed. Such collaborations will facilitate the faster and more effective development of AI systems. The use of AI systems in the healthcare sector requires continuous monitoring and evaluation processes. The performance, accuracy, and reliability of algorithms should be regularly monitored and evaluated, and necessary updates and improvements should be made to the systems accordingly. Moreover, the data obtained from the use of AI systems will be a valuable data source for future studies and developments. These data should be used to identify the weaknesses of AI systems and make improvements in these areas. Comparing the behavioral characteristics between experts and AIs enables a better understanding of the training and adaptation processes of AIs. Factors such as flexibility, experience, and intuition in the decision-making processes of experts stand out as areas where AI systems can be further developed. On the other hand, the ability of AIs to analyze large datasets quickly and accurately can reduce the workload of experts and speed up diagnostic processes. This comparison could allow for the delivery of more effective and efficient healthcare services by combining the strengths of both parties.

The results of this study demonstrate how AI technologies behave in medical diagnosis and evaluation processes and the crucial role they can play. However, it should not be forgotten that the human factor should not be overlooked for these technologies to be used effectively and safely. Future research and developments will be essential steps to promote the wider adoption and use of AI systems in the healthcare sector.

In the next decade, the collaboration between humans and AI is expected to evolve further. This cooperation, which is anticipated to bring significant changes and innovations in many fields, will redefine the way we work. This collaboration will impact many areas, from business to daily life, with the rapid advancement of technology and the increasing integration of AI. It is expected that AI will take on routine and repetitive tasks, allowing humans to focus on more creative and strategic tasks. Additionally, new job areas and professions are expected to emerge, particularly in AI development and management, contributing to employment growth. New training and skill development programs will need to be created to ensure that workers can work harmoniously and effectively with AI. AI is expected to be used

more effectively in the early diagnosis and treatment of diseases, which is anticipated to result in significant progress in medical imaging and genetic analysis, along with increased efficiency and reduced costs. It is expected that the developing collaboration between humans and AI will lead to the development of personalized treatment methods in healthcare services. Especially when the decision-making behaviors of AI are more clearly defined, doctors and healthcare personnel will be able to make faster and more accurate decisions with AI-supported systems. AI is expected to analyze student performance and provide personalized education programs, while virtual classrooms and online learning platforms are anticipated to become more effective with AI-supported content. By optimizing educational materials and teaching methods, AI is expected to increase the efficiency of education, thereby reducing the duration and processes of education. The key to all these developments will lie in the concept of bias, which defines and guides AI's behavior and represents both its strongest and weakest trait.

REFERENCES

[1] Turing, A. M.: Computing Machinery and Intelligence. Mind, Vol. 59, 1950, No. 236, pp. 433–460, doi: 10.1093/mind/LIX.236.433.

[2] Russell, S. J.—Norvig, P.: Artificial Intelligence: A Modern Approach. 3rd Edition. Prentice Hall, 2009.

[3] Homer: The Iliad of Homer. University of Chicago Press, 1962.

[4] Boden, M. A.: Artificial Intelligence: A Very Short Introduction. Oxford University Press, 2018, doi: 10.1093/actrade/9780199602919.001.0001.

[5] Nilsson, N. J.: Artificial Intelligence: A New Synthesis. Morgan Kaufmann, 1998, doi: 10.1016/C2009-0-27773-7.

[6] Poole, D. L.—Mackworth, A. K.: Artificial Intelligence: Foundations of Computational Agents. 2nd Edition. Cambridge University Press, 2017, doi: 10.1017/9781108164085.

[7] Över Özçelik, T.—Selvi, İ. H.—Yilmaz Yalçiner, A.—Zeren, M. T.: Industry 4.0, Smart Factories and Effects on Business. In: Torkul, O., Tunacan, T. (Eds.): Knowledge Management and Digital Transformation Power. Efe Akademi Publishing, 2022, pp. 169–190.

[8] Zeren, M. T.—Aytulun, S. K.—Kirelli, Y.: Comparison of SSD and Faster R-CNN Algorithms to Detect the Airports with Data Set Obtained from Unmanned Aerial Vehicles and Satellite Images. European Journal of Science and Technology (Avrupa Bilim ve Teknoloji Dergisi), 2020, No. 19, pp. 643–658, doi: 10.31590/ejosat.742789.

[9] Deep Blue (Chess Computer). Wikipedia, The Free Encyclopedia, 2023, `https://en.wikipedia.org/wiki/Deep_Blue_(chess_computer)` [accessed 2023-05-20] (Version Dated 20 May 2023).

[10] IBM Watson. International Business Machines (IBM), `https://www.ibm.com/watson`.

[11] AlphaGo: The Story So Far. Google DeepMind, https://deepmind.google/research/alphago/.

[12] Spice, B.: Carnegie Mellon Artificial Intelligence Beats Top Poker Pros. Carnegie Mellon University News, 2017, https://www.cmu.edu/news/stories/archives/2017/january/ai-beats-poker-pros.html.

[13] Brockman, G.—Dennison, C.—Zhang, S.—Pachocki, J.—Petrov, M. et al.: OpenAI Five. OpenAI, 2018, https://openai.com/research/openai-five.

[14] IBM Project Debater. IBM Research, 2019, https://www.research.ibm.com/artificial-intelligence.

[15] GPT-3 Powers the Next Generation of Apps. OpenAI, 2021, https://openai.com/index/gpt-3-apps/.

[16] Jumper, J.—Evans, R.—Pritzel, A.—Green, T.—Figurnov, M. et al.: Highly Accurate Protein Structure Prediction with AlphaFold. Nature, Vol. 596, 2021, No. 7873, pp. 583–589, doi: 10.1038/s41586-021-03819-2.

[17] Zheng, Y.—Sun, X.—Feng, B.—Kang, K.—Yang, Y.—Zhao, A.—Wu, Y.: Rare and Complex Diseases in Focus: ChatGPT's Role in Improving Diagnosis and Treatment. Frontiers in Artificial Intelligence, Vol. 7, 2024, Art. No. 1338433, doi: 10.3389/frai.2024.1338433.

[18] Bhasuran, B.—Schmolly, K.—Kapoor, Y.—Jayakumar, N. L.—Doan, R.—Amin, J. et al.: Reducing Diagnostic Delays in Acute Hepatic Porphyria Using Health Records Data and Machine Learning. Journal of the American Medical Informatics Association, Vol. 32, 2025, No. 1, pp. 63–70, doi: 10.1093/jamia/ocae141.

[19] Cheng, J.—Novati, G.—Pan, J.—Bycroft, C.—Žemgulytė, A.—Applebaum, T. et al.: Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense. Science, Vol. 381, 2023, No. 6664, Art. No. eadg7492, doi: 10.1126/science.adg7492.

[20] Hall, J. E.—Hall, M. E.: Guyton and Hall Textbook of Medical Physiology. 14th Edition. Elsevier Health Sciences, 2020.

[21] Upadhyay, R. S.—Tanwar, P.: A Review on Bone Fracture Detection Techniques Using Image Processing. 2019 International Conference on Intelligent Computing and Control Systems (ICCS), IEEE, 2019, pp. 287–292, doi: 10.1109/ICCS45141.2019.9065874.

[22] Haleem, S.—Lutchman, L.—Mayahi, R.—Grice, J. E.—Parker, M. J.: Mortality Following Hip Fracture: Trends and Geographical Variations over the Last 40 Years. Injury, Vol. 39, 2008, No. 10, pp. 1157–1163, doi: 10.1016/j.injury.2008.03.022.

[23] Court-Brown, C. M.—Heckman, J. D.—McKee, M.—McQueen, M. M.—Ricci, W.—Tornetta, P.: Rockwood and Green's Fractures in Adults. 8th Edition. Lippincott Williams & Wilkins, 2014.

[24] Filipov, O.: Epidemiology and Social Burden of the Femoral Neck Fractures. Journal of IMAB, Vol. 20, 2014, No. 4, pp. 516–518, doi: 10.5272/jimab.2014204.516.

[25] Gallagher, J. C.—Melton, L. J.—Riggs, B. L.—Bergstrath, E.: Epidemiology of Fractures of the Proximal Femur in Rochester, Minnesota. Clinical Orthopaedics and Related Research, Vol. 150, 1980, pp. 163–171.

[26] Berglund-Rödén, M.—Swierstra, B. A.—Wingstrand, H.—Thorngren, K. G.: Prospective Comparison of Hip Fracture Treatment: 856 Cases Followed for 4 Months in The Netherlands and Sweden. Acta Orthopaedica Scandinavica, Vol. 65, 1994, No. 3, pp. 287–294, doi: 10.3109/17453679408995455.

[27] Korteling, J. E.—van de Boer-Visschedijk, G. C.—Blankendaal, R. A.—Boonekamp, R. C.—Eikelboom, A. R.: Human- Versus Artificial Intelligence. Frontiers in Artificial Intelligence, Vol. 4, 2021, Art. No. 622364, doi: 10.3389/frai.2021.622364.

[28] Triberti, S.—Durosini, I.—Lin, J.—La Torre, D.—Ruiz Galán, M.: Editorial: On the "Human" in Human–Artificial Intelligence Interaction. Frontiers in Psychology, Vol. 12, 2021, Art. No. 808995, doi: 10.3389/fpsyg.2021.808995.

[29] Thirunavukarasu, A. J.—Mahmood, S.—Malem, A.—Foster, W. P.—Sanghera, R.—Hassan, R. et al.: Large Language Models Approach Expert-Level Clinical Knowledge and Reasoning in Ophthalmology: A Head-to-Head Cross-Sectional Study. PLOS Digital Health, Vol. 3, 2024, No. 4, Art. No. e0000341, doi: 10.1371/journal.pdig.0000341.

[30] Ippolito, D.—Maino, C.—Gandola, D.—Franco, P. N.—Miron, R.—Barbu, V.—Bologna, M.—Corso, R.—Breaban, M. E.: Artificial Intelligence Applied to Chest X-Ray: A Reliable Tool to Assess the Differential Diagnosis of Lung Pneumonia in the Emergency Department. Diseases, Vol. 11, 2023, No. 4, Art. No. 171, doi: 10.3390/diseases11040171.

[31] Rahwan, I.—Cebrian, M.—Obradovich, N.—Bongard, J.—Bonnefon, J. F.—Breazeal, C. et al.: Machine Behaviour. Nature, Vol. 568, 2019, No. 7753, pp. 477–486, doi: 10.1038/s41586-019-1138-y.

[32] Pedersen, T.—Johansen, C.: Behavioural Artificial Intelligence: An Agenda for Systematic Empirical Studies of Artificial Inference. AI & Society, Vol. 35, 2020, No. 3, pp. 519–532, doi: 10.1007/s00146-019-00928-5.

[33] Dietvorst, B. J.—Simmons, J. P.—Massey, C.: Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. Journal of Experimental Psychology: General, Vol. 144, 2015, No. 1, pp. 114–126, doi: 10.1037/xge0000033.

[34] Cummings, M. L.: Man Versus Machine or Man + Machine? IEEE Intelligent Systems, Vol. 29, 2014, No. 5, pp. 62–69, doi: 10.1109/MIS.2014.87.

[35] de Visser, E. J.—Parasuraman, R.: Adaptive Aiding of Human–Robot Teaming: Effects of Imperfect Automation on Performance, Trust, and Workload. Journal of Cognitive Engineering and Decision Making, Vol. 5, 2011, No. 2, pp. 209–231, doi: 10.1177/1555343411410160.

[36] Silver, D.—Huang, A.—Maddison, C. J.—Guez, A.—Sifre, L.—van den Driessche, G. et al.: Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature, Vol. 529, 2016, No. 7587, pp. 484–489, doi: 10.1038/nature16961.

[37] McDuff, D.—El Kaliouby, R.—Senechal, T.—Amr, M.—Cohn, J. F.—Picard, R.: Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected "In-the-Wild". 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 881–888, doi:

10.1109/CVPRW.2013.130.

[38] Hancock, P. A.—Billings, D. R.—Schaefer, K. E.—Chen, J. Y. C.—de Visser, E. J.—Parasuraman, R.: A Meta-Analysis of Factors Affecting Trust in Human–Robot Interaction. Human Factors, Vol. 53, 2011, No. 5, pp. 517–527, doi: 10.1177/0018720811417254.

[39] Nass, C.—Fogg, B. J.—Moon, Y.: Can Computers Be Teammates? International Journal of Human-Computer Studies, Vol. 45, 1996, No. 6, pp. 669–678, doi: 10.1006/ijhc.1996.0073.

[40] Bryson, J. J.—Winfield, A. F. T.: Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. Computer, Vol. 50, 2017, No. 5, pp. 116–119, doi: 10.1109/MC.2017.154.

[41] Yeomans, M.—Shah, A.—Mullainathan, S.—Kleinberg, J.: Making Sense of Recommendations. Journal of Behavioral Decision Making, Vol. 32, 2019, No. 4, pp. 403–414, doi: 10.1002/bdm.2118.

[42] Luckin, R.—Holmes, W.—Griffiths, M.—Forcier, L. B.: Intelligence Unleashed: An Argument for AI in Education. Pearson Education, 2016, https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/open-ideas/Intelligence-Unleashed-v15-Web.pdf.

[43] Colton, S.—Wiggins, G. A.: Computational Creativity: The Final Frontier? In: De Raedt, L., Bessiere, C., Dubois, D., Doherty, P., Frasconi, P., Heintz, F., Lucas, P. (Eds.): ECAI 2012: 20th European Conference on Artificial Intelligence. IOS Press, Frontiers in Artificial Intelligence and Applications, Vol. 242, 2012, pp. 21–26, doi: 10.3233/978-1-61499-098-7-21.

[44] Floridi, L.—Cowls, J.: A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, Vol. 1, 2019, No. 1, doi: 10.1162/99608f92.8cd550d1.

[45] Esteva, A.—Kuprel, B.—Novoa, R. A.—Ko, J.—Swetter, S. M.—Blau, H. M.—Thrun, S.: Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. Nature, Vol. 542, 2017, No. 7639, pp. 115–118, doi: 10.1038/nature21056.

[46] Lakhani, P.—Sundaram, B.: Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology, Vol. 284, 2017, No. 2, pp. 574–582, doi: 10.1148/radiol.2017162326.

[47] Rajpurkar, P.—Irvin, J.—Ball, R. L.—Zhu, K.—Yang, B.—Mehta, H. et al.: Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practising Radiologists. PLoS Medicine, Vol. 15, 2018, No. 11, Art. No. e1002686, doi: 10.1371/journal.pmed.1002686.

[48] Gulshan, V.—Peng, L.—Coram, M.—Stumpe, M. C.—Wu, D. et al.: Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, Vol. 316, 2016, No. 22, pp. 2402–2410, doi: 10.1001/jama.2016.17216.

[49] Liu, X.—Faes, L.—Kale, A. U.—Wagner, S. K.—Fu, D. J.—Bruynseels, A. et al.: A Comparison of Deep Learning Performance Against Health-Care Professionals in Detecting Diseases from Medical Imaging: A Systematic Review and

Meta-Analysis. The Lancet Digital Health, Vol. 1, 2019, No. 6, pp. e271–e297, doi: 10.1016/S2589-7500(19)30123-2.

[50] Zeren, M. T.—Arslankaya, S.—Altuntaş, Y.—Cam, N.—Kirelli, Y.—Özdemir, M. H.: Doctors Versus YOLO: Comparison Between YOLO Algorithm, Orthopedic and Traumatology Resident Doctors and General Practitioners on Detection of Proximal Femoral Fractures on X-Ray Images with Multi Methods. International Journal on Artificial Intelligence Tools, Vol. 33, 2024, No. 1, Art. No. 2350056, doi: 10.1142/S0218213023500562.

[51] Shen, J.—Zhang, C. J. P.—Jiang, B.—Chen, J.—Song, J.—Liu, Z.—He, Z.—Wong, S. Y.—Fang, P. H.—Ming, W. K.: Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. JMIR Medical Informatics, Vol. 7, 2019, No. 3, Art. No. e10010, doi: 10.2196/10010.

[52] Baker, A.—Perov, Y.—Middleton, K.—Baxter, J.—Mullarkey, D.—Sangar, D.—Butt, M.—DoRosario, A.—Johri, S.: A Comparison of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis. Frontiers in Artificial Intelligence, Vol. 3, 2020, Art. No. 543405, doi: 10.3389/frai.2020.543405.

[53] Bhuva, A. N.—Bai, W.—Lau, C.—Davies, R. H.—Ye, Y.—Bulluck, H. et al.: A Multicenter, Scan–Rescan, Human and Machine Learning CMR Study to Test Generalizability and Precision in Imaging Biomarker Analysis. Circulation: Cardiovascular Imaging, Vol. 12, 2019, No. 10, Art. No. e009214, doi: 10.1161/CIRCIMAGING.119.009214.

[54] Haenssle, H. A.—Fink, C.—Schneiderbaue, R.—Toberer, F.—Buhl, T.—Blum, A. et al.: Man Against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists. Annals of Oncology, Vol. 29, 2018, No. 8, pp. 1836–1842, doi: 10.1093/annonc/mdy166.

[55] Niebel, B. W.—Freivalds, A.: Niebel's Methods, Standards, and Work Design. 13th Edition. McGraw-Hill Education, 2013.

[56] Zandin, K. B.: Maynard's Industrial Engineering Handbook. 5th Edition. McGraw-Hill Professional, 2001.

[57] Barnes, R. M.: Motion and Time Study: Design and Measurement of Work. 7th Edition. John Wiley & Sons, 1980.

[58] International Labour Office: Introduction to Work Study. International Labour Organization, 1992.

[59] Montgomery, D. C.—Runger, G. C.: Applied Statistics and Probability for Engineers. 5th Edition. John Wiley & Sons, 2010.

[60] Ross, S. M.: Introductory Statistics. 4th Edition. Academic Press, 2017.

[61] McDonald, J. H.: Handbook of Biological Statistics. 3rd Edition. Sparky House Publishing, 2014.

[62] Kayri, M.: The Multiple Comparison (Post-Hoc) Techniques to Determine the Difference Between Groups in Researches. Fırat University Journal of Social Science, Vol. 19, 2009, No. 1, pp. 51–64 (in Turkish).

[63] Tukey, J. W.: Comparing Individual Means in the Analysis of Variance. Biometrics, Vol. 5, 1949, No. 2, pp. 99–114, doi: 10.2307/3001913.

[64] Tabachnick, B. G.—Fidell, L. S.: Using Multivariate Statistics. 6th Edition. Pearson, 2013.

[65] George, D.—Mallery, P.: SPSS for Windows Step by Step: A Simple Guide and Reference. 11th Edition. Pearson Education India, 2011.

Yusuf Altuntaş is a Consultant Orthopaedic Surgeon at the Department of Orthopaedics and Traumatology, Sisli Hamidiye Etfal Training and Research Hospital, University of Health Sciences, Türkiye. His clinical and research interests mainly focus on sports surgery, particularly knee surgery and rotator cuff disorders, as well as arthroplasty. He also has interests in spine surgery and orthopaedic trauma. Between 2013 and 2014, he gained international experience at the University of Arizona, Tucson, where he collaborated with the Department of Neurosurgery in the field of vertebral surgery and spinal tumors. He completed his medical education and orthopaedic residency training at Sisli Hamidiye Etfal Training and Research Hospital, University of Health Sciences, and obtained national board certification in Orthopaedics and Traumatology in 2024. In addition to his clinical practice, he has a growing interest in the applications of artificial intelligence in medicine and orthopaedic research. He is also a member of the Editorial Board of Annals of Orthopedics and Traumatology (AOT).

Muhammed Taha Zeren is Ph.D. industrial engineer specializing in artificial intelligence, machine learning, deep learning, and computer vision. He holds his B.Sc. in industrial engineering from Kocaeli University, M.Sc. in industrial engineering from Beykent University, and Ph.D. in industrial engineering from Sakarya University. He has held managerial roles in leading organizations, including at Merkezi Kayıt Kuruluşu A.Ş. (MKK), Doctors Worldwide Türkiye, LC Waikiki, Turkish Aerospace Industries, TRT World, and Ford Motor Company. Since May 2025, he has been working as Project Manager responsible for R & D Projects at MKK. He has also given lectures at universities, particularly İstinye University, on deep learning, digital transformation, machine learning, computer vision, and project management. His academic publications cover medical image analysis, object detection, fraud detection, Industry 4.0, digital transformation, and financial technologies. He combines academic expertise with extensive experience in R & D, IT, project and program management, operational excellence, and strategic execution.

Şebnem Özdemir is Associated Professor and holds her Bachelor's degree in mathematics from Yıldız Technical University, two Master's degrees in mathematics education and informatics, and a Ph.D. in informatics from the Istanbul University. She conducted postdoctoral research at Worcester Polytechnic Institute and MIT CSAIL in the United States. She currently serves as Head of the Department of Management Information Systems and leads the Department of Data Science at İstinye University. Since March 2025, she has also served as Chief Artificial Intelligence Officer at MLP CARE. She is the founder of Usight Software and Artificial Intelligence Technologies Inc., providing AI-driven solutions and digital transformation consultancy. She has received several awards for her contributions to artificial intelligence and data science. She has participated in numerous national and international projects as a principal investigator, consultant, and researcher. She is also the author of scientific publications and book chapters and holds several national and international patents.