



IJAMEC

**International Journal of
Applied Methods in
Electronics and Computers**

Vol 14 (1)

2026

Contents	Pages
<i>Research Articles</i>	
Early Diagnosis of Tomato Diseases Using Machine Learning to Reduce Pesticide Application Nasir Cetin, Ziya Ozcelik, Yavuz Selim Taspinar	1-13
Simulation-Based Evaluation of LiDAR–Photogrammetry Fusion via NeRF Reconstruction and ICP Registration in Urban Scenes Rytis Maskeliūnas, Sarmad Maqsood, Ahmad Qurthobi, Irfan Abbas, Mantas Vaškevičius, Julius Gelšvartas	14-19
Classification of Orthopedic Diseases Using Biomechanical Properties with Machine Learning Methods Mehmet Cüneyt Özbacı, Turgay Tugay BİLGİN	20-33
Deep Learning-Based Lung Cancer Classification and Grad-Cam++ With Lime-Supported Explainability Analysis Yasin DOĞANŞAH, İnayet Hakki CİZMECI, Murat KOKLU	34-48
Domain-Specific Fine-Tuning of LLMs for Elevator Emergency Response Systems Busra Onal, Muhammet Fatih Aslan, Akif Durdu	49-58

*Research Article***Early Diagnosis of Tomato Diseases Using Machine Learning to Reduce Pesticide Application****Nasir Cetin ^{a,*}, Ziya Ozcelik ^b, Yavuz Selim Taspinar ^c**^a *Integra Hydraulic, Konya, Türkiye.*^b *Selçuk University, Faculty of Technology, Department of Mechatronic Engineering, Konya, Türkiye.*^c *Selçuk University, Faculty of Technology, Department of Mechatronic Engineering, Konya, Türkiye.*

ARTICLE INFO

Article history:

Received 26 December 2025

Accepted 30 March 2026

*Keywords:*Classification,
Machine learning,
Organic farming,
Pesticides,
Tomato diseases

ABSTRACT

Pesticides, which are widely used today to increase yield and prevent diseases in agricultural products, have negative effects on human health and the environment. Due to these effects, efforts to restrict pesticide use gained momentum with environmental protection movements starting in the 1970s. Strict controls and licensing processes have been introduced for pesticide production to prevent their unconscious use. The aim of this study is to classify diseases observed in tomato plants, thereby preparing a scientific groundwork for reducing pesticide use. For this purpose, a dataset comprising a total of 13,883 records of tomato plants belonging to six different classes was utilized. For the classification of the data, the following Machine Learning methods were employed: Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Logistic Regression (LR) algorithms. Model performance was evaluated using metrics such as accuracy, precision, recall, and F1 Score, along with the confusion matrix and ROC analysis. The classification success rates obtained from model training and testing were determined as: ANN 100.0%, KNN 99.8%, SVM 99.9%, and LR 99.9%. The results indicate that all models achieved high and very similar success rates. In conclusion, it is expected that all classification models used can be successfully utilized for the reliable diagnosis of tomato plant diseases, and this capability is projected to make significant contributions to organic farming applications.

This is an open access article under the CC BY-SA 4.0 license.
(<https://creativecommons.org/licenses/by-sa/4.0/>)

1. Introduction

Pesticides are chemical substances or mixtures of substances used to control or prevent damage from insects, weeds, microorganisms and pests that damage agricultural products or animal food products during production, harvesting or display [1]. Pesticides are not applied directly to plants or soil. Because they are toxic substances, they are used by mixing them with some auxiliary substances (solid, liquid) in a way that minimizes harm to human and environmental health against pests. The pesticide in this mixture is called the "active substance" or "inactive ingredient". Not every toxic substance can be used and named as a pesticide. For a toxic

substance to be considered a pesticide, it must be biologically active, effective, sufficiently stable, safe for users, third parties and consumers, safe for livestock, not harmful to wildlife and beneficial organisms, acceptable for the environment and have characteristics that will not cause problems in trade. Many studies are being conducted on this subject in the literature. Some of those related to this study are given in order [1].

In one study, Tiryaki et al. conducted the necessary research by investigating the benefits and harms of pesticides using data on average pesticide consumption per unit area in some European Union (EU) countries. Their contribution to the literature was to compare pesticide use in Turkey, EU countries, and the world, explaining the

* Corresponding author. E-mail address: nasr.cetin@gmail.com
DOI: 10.58190/ijamec.2026.161

advantages and disadvantages of pesticide use and the behavior of pesticides in the agricultural ecosystem. They also addressed issues such as the safe and effective application of pesticides, possible residues in products, and factors affecting these residues. In addition, they compared the situation of countries that received residue alerts published in the EU rapid alert system [2].

Özcan and Tongur analyzed the toxic effects of insecticides and herbicides using the *Lepidium sativum* toxicity test method, determined that herbicides were more toxic than insecticides, and found that the harm of some banned pesticides reached high levels [3].

In their survey study with rose oil producers in Isparta, Arıcı et al. observed that a significant portion of the producers did not adequately consider the environmental risks when applying pesticides, but a large portion of them complied with the waiting periods for the pesticides [4].

In one of their studies, Yüzbaşıoğlu and Topkaya conducted a study on pesticide use with 95 plant producers randomly selected from villages in the central district of Tokat province. As a result of the interviews, it was observed that the producers were within the working population and had not received any education beyond compulsory education. It was determined that the producers were influenced by pesticide dealers in their pesticide use. It can be said that they did not take any protective measures during agricultural spraying. It was observed that the producers, who thought that pesticide use did not harm the product, spent very high prices financially. It was observed that they were unaware of the environmental damage caused by the disposal of the waste from the pesticides they used. In conclusion, it can be said that the producers have very little knowledge about pesticide use and its aftermath. As a contribution to the literature, they determined that there is a need for relevant institutions and organizations to organize seminars on pesticide application and post-application information [5].

Polat stated that global pesticide use has reached 3.5 billion kg annually in the last 20 years, and in Türkiye, use increased by 11% compared to the previous year, reaching 60 thousand tons. The study emphasized that chemical control alone is not sustainable due to risks such as environmental toxicity and resistance development. In conclusion, the author argues that pesticides are necessary but should be applied in a controlled manner, integrated with alternative control methods, in order to prevent ecological damage [6].

In a study, Özeran and Taşçı examined pesticide usage amounts in Turkey based on provinces, regions, and pesticide groups, using the most recent data available for 2020. Based on this data, the province with the highest pesticide usage in Turkey is Antalya, while the province with the lowest usage is Ardahan. Pesticides are used most in the Mediterranean Region due to the diversity of crops grown there, while the region with the lowest use is the

Black Sea Region. The most widely used pesticide groups in Turkey in 2020 were fungicides (38.4%), herbicides (27.4%), and insecticides (23.0%), respectively. When examining the top three provinces with the most intensive pesticide use, it was found that the most widely used pesticide group was fungicides in Antalya (38.3%) and Manisa (81.0%), and insecticides in Adana (65.2%). Chemical control is the most preferred agricultural control method in Turkey for combating diseases, pests, and weeds to ensure sustainable production. Although pesticide use seems inevitable, it has been stated that raising producer awareness and promoting integrated control would have positive effects on human and environmental health by reducing usage [7].

In a study, Yılmaz et al. aimed to analyze changes in pesticide use by utilizing data from the United Nations Food and Agriculture Organization (FAO) and data on pesticides and pesticide groups used worldwide, in EU countries, and in Turkey between 2000 and 2021, as found in articles and reports on the subject. According to the findings of the study, as of 2021, the global average pesticide use was 1.07 kg ha⁻¹, the EU average was 3.20 kg ha⁻¹, and the Turkish average was 2.26 kg ha⁻¹. Accordingly, the amount of pesticides used per hectare in Turkey is lower than the EU average but higher than the global average. While herbicides have the highest share of pesticide use per unit area worldwide (21.92%), fungicides and bactericides have the highest share in the EU (43.18%) and Turkey (36.06%). While the European Union has implemented comprehensive legal regulations to reduce pesticide use and ensure environmental sustainability, regulations on pesticide use in Turkey remain limited. Despite the inadequacy of legal regulations on pesticide use in Turkey, it can be said that there have been significant efforts in recent years to promote sustainable agricultural practices. According to the literature, Turkey should develop more specific legal regulations, policies, and strategies that reduce the risks associated with pesticide use, taking into account the EU harmonization process, and implement the EU's structurally focused and comprehensive legal regulations. They have contributed by suggesting the implementation of more effective agricultural environmental policy measures, such as taxation and subsidies, to develop alternative disease and pest control methods and ensure farmer adaptation for sustainable agricultural production in Turkey [8].

In a study, Ayyıldız utilized FMOLS-DOLS models to determine the long-term effects of pesticides using chemical pesticide, environmental, and economic indicator data from 1990 to 2021. According to the findings, fungicide and herbicide use had a significant effect on the crop production index and agricultural greenhouse gas emissions, while insecticide use was found to be statistically insignificant. The results suggest that pesticide use contributes significantly to productivity but

causes increasing environmental damage over time. In the literature, the author considers it important to establish legal regulations aimed at reducing pesticide use and ensuring its correct use, and to activate publication activities. He has contributed by stating that priority should be given to R&D studies for the widespread use and improved effectiveness of biopesticides as an alternative to chemical pesticide use [9].

A study by Tıraş found that as of 2020, the top six countries with the highest agricultural production globally (China, India, Brazil, the United States, Indonesia, and Russia) using panel data from 1995 to 2019, estimated the effect of fertilizer and pesticide use on agricultural Gross Domestic Product (GDP) using panel econometric methods. It stated that the effect of fertilizer use on agricultural GDP was significant, but the effect of pesticide use was statistically insignificant [10].

Doğan and Karpuzcu simulated the effects of structural (filter strips) and non-structural (use reduction) Best Management Practices (BMPs) on pesticide load in a water basin in Istanbul using a SWAT model. In the study, it was predicted that 25-meter-wide filter strips provided 95% retention and that the basin would be cleaned in 3-4 years if pesticide use were completely eliminated. The authors emphasized that mathematical models play a critical role in decision support processes, but the model sensitivity should be increased with additional parameters (Koc, land slope, etc.) [11].

In a study, Aydoğan and Baran used data obtained from surveys of 94 tobacco farmers in the Çelikhan district of Adıyaman province and determined that 44.0% of farmers had a low level of awareness of agricultural pest control, 45.0% had a medium level, and 11.0% had a high level. Furthermore, differences were observed in the awareness levels of farmers before, during, and after application, and it was found that farmers adhered more strictly to the rules before agricultural pest control. The results were used in the field of agricultural activities. Literature In the agricultural education and extension activities planned in the region, supporting educational content that conveys the negative effects of agricultural pesticides to farmers with visuals and strictly monitoring prescription-based pesticide use may help solve the problem. On the other hand, the direct consumption of Çelikhan tobacco leaves has contributed to the emergence of the residue problem and the need for further studies on pesticide residues in tobacco and their effects on human health [12].

In a study, Hayran et al. used data collected from 68 randomly selected farmers to evaluate pesticide use, analyze farmers' risk perception regarding pesticide use, and identify its determinants using descriptive statistics and logistic regression analysis methods. Based on these methods, it was determined that farmers' risk perceptions were at a moderate level and that these perceptions were influenced by variables such as education, experience,

internet use, non-agricultural income, and greenhouse area. This is one of the limited studies in the literature examining greenhouse farmers' risk perceptions regarding pesticide use, and the authors have contributed to this field [13].

In one study, Doğan and Karaaslan examined the production practices of grain producers benefiting from the ÇATAK program in the Seyfe Gölü region of Kırşehir province from an environmental perspective. They found that producers relied on their experience in the use of chemical fertilizers and pesticides and focused on cost rather than environmental risk, recommending that local actors keep their information activities dynamic [14].

In a study, Tözün and Akar conducted a systematic review of articles published in the national literature in Turkey and found a total of 35 records that met the inclusion criteria using search data from Izmir Katip Çelebi University Electronic Library resources and Google Scholar. It was determined that 16 (45.7%) of these contained pesticide residues and exceeded the Maximum Residue Levels (MRL); 10 (28.6%) contained pesticide residues but did not exceed the MRL values. Although the food samples studied varied, pesticide residues were studied extensively and found in grapes, strawberries, citrus fruits, and some commonly consumed vegetables. Milk and honey are notable among the foods in which no pesticide residues were found. They contributed to the literature with recommendations such as reviewing the legislation on the subject and introducing deterrent penalties, as well as increasing educational efforts on the subject for public and occupational health [15].

In a study, Gedikli et al. 2011 used a simple random sampling method in the districts of Alaçam, Bafra, and Terme in the province of Samsun during the 2011 production year to survey 99 farmers in 15 villages regarding their environmental awareness in the use of agricultural pesticides in rice, corn, and wheat production. In addition to these data, data from local and foreign literature and state institutes and organizations were used. According to the research results, although agricultural institutions are not at a sufficient level, they are visited by farmers. Farmers consult with pesticide dealers and agricultural engineers in the selection of chemical pesticides (48.8%) and dosage adjustment (33.34%), and they obtain pesticides when they start to see damage (70.70%). In addition, 67.67% of producers dispose of pesticide packaging by burning it after spraying, 26.26% bury it in the ground, and 6.06% dispose of it in the trash by bagging it. The literature on harmful species (disease, pest, and weed) and their environmental relationships, and ensuring that the population of harmful species is kept below the level of economic damage. Thus, by researching and applying alternative methods to chemical control, the approach of minimizing pesticide use will be adopted, thereby contributing to the significant protection of both

the natural environment and human health [16].

The primary objective of this study is to diagnose diseases and weeds affecting tomato plants with high accuracy and at an early stage using deep learning architectures and machine learning algorithms as an alternative to traditional methods. The rapid and accurate detection of plant diseases aims to reduce environmental impacts and maintain production efficiency by enabling the application of pesticides only when necessary and in the correct amounts.

The contributions of this study to the literature and practice can be summarized as follows:

- **Early Diagnosis with Deep Learning:** Using the SqueezeNet deep learning architecture, early diagnosis was made possible through visual data containing healthy and diseased states of tomato plants.

- **Multi-Classification Comparison:** Classical machine learning algorithms such as Artificial Neural Networks (ANN), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), and Logistic Regression (LR) were systematically applied on deep features obtained from SqueezeNet, and their performances were compared. This approach demonstrates the potential for integrating transfer learning with classical ML.

- **Application on a Large Dataset:** The study evaluated the performance of multi-class detection using a six-class (diseased and healthy) tomato leaf dataset consisting of 13,883 different photographs.

- **Weed Segregation:** By detecting healthy plants, the study indirectly demonstrates the potential for segregating harmful weeds growing alongside them, paving the way for precision farming applications.

The remainder of the study is organized as follows: the second section, the material and methods section, presents the dataset used in the study, the methods employed, and the performance metrics. The third section presents the experimental results obtained from the study and the discussions. The fourth section presents the results, contributions, limitations, and areas of application of the study.

2. Material and Methods

2.1. Crop Disease Detection Dataset (CDDD)

The dataset used in the study contains a total of 13,883 images across 6 classes. There are 2,127 images in the Bacterial Spot class, 2,400 in the Early Blight class, 2,407 in the Healthy class, 2,314 in the Late Blight class, 2,184 in the Septoria Leaf Spot class, and 2,451 in the Yellow Leaf Curl Virus class. Each image is 256x256 pixels in size. Figure 1 shows sample images from the dataset [17].

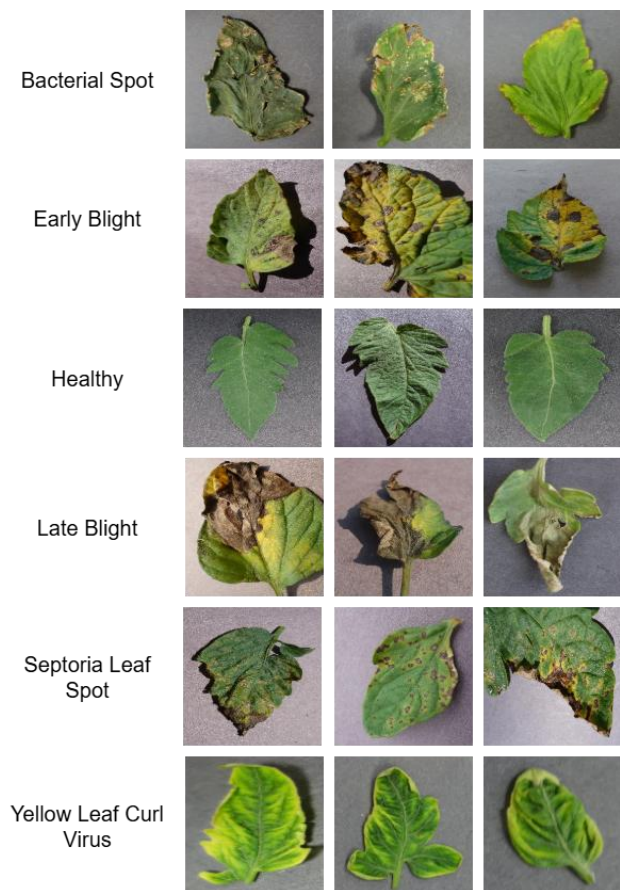


Figure 1. Sample images by class in CDDD

Bacterial Spot disease begins as water-soaked spots; the spots increase in number and may merge to form necrotic areas on the leaves, giving them a burnt appearance. Leaf drop can cause severe defoliation and make the fruit susceptible to sunscald. Mature spots have an oily appearance and may appear transparent when held up to the light. The centers of the spots dry out and fall off the leaf; burnt leaves usually remain attached to the plant and give it a burnt appearance; fruit infections start as slightly raised blisters; spots may have a slight halo that disappears over time; lesions on the fruit may have a raised edge and a sunken center, giving the fruit a pitted appearance [18].

Early Blight disease symptoms begin as oval-shaped lesions with a yellow chlorotic zone along the lesions; concentric leaf remnants may be seen on fused leaves; leaf structure between veins is provided; severe blight can cause complete collapse of the leaves; the disease severely burns the leaves and leads to loss of yield; tomato stems can become infected with the fungus and lead to Alternaria stem canker; the first signs of stem cankers are changes of dark brown areas on the stem; stem cankers can grow to encompass the entire stem and cause the entire main center; brown lines may be present on top of the canker and below the veins; fruit symptoms include small black v-shaped lesions on the shoulders of the fruit (the disease is also called black shoulder); lesions are also resistant as dark spots with a concentric ring pattern on the fruits; fruit pieces may be visible in the field or may develop during

the presentation of the fruit to the market; lesions may have a velvety appearance due to sporulation of the fungus [18].

Late Blight disease affects all above-ground parts of the tomato plant; the first signs of the disease appear as waterlogged green to black areas on the leaves and rapidly turn into brown lesions; in wet weather, fluffy white fungal growth may be seen in infected areas and on the undersides of the leaves; as the disease progresses, the leaves shrivel and turn brown and the whole plant may die; fruit lesions begin as irregularly shaped waterlogged areas and turn into oily spots; the entire fruit may become infected and a white fuzzy growth may be seen in wet weather [18].

Symptoms of Septoria Leaf Spot disease can appear at any stage of tomato development and may begin as small, water-soaked spots or circular gray-white spots on the underside of older leaves; the spots have a gray center and a dark edge and may sink inwards; fungal fruiting bodies may be visible as small black dots in the center of the spot; spots may also be seen on stems, fruit calyxes and flowers [18].

In Yellow Leaf Curl Virus disease symptoms, infected leaves shrink, curl upwards, appear wrinkled, and yellowing is seen in the veins and leaf edges. Internodes shorten and the whole plant appears stunted and bushy. The whole plant stands upright and only upright growth is observed. Flowers may not develop and may fall off [18].

2.2. Convolutional Neural Network (CNN)

Deep learning is a subfield of artificial intelligence. It involves developing artificial neural networks to perform mathematical and statistical calculations and process data related to a problem. The processed data is then transformed into predictions or results at the final layer of the neural network. The success of the prediction is measured by evaluating these results using various metrics. Training artificial neural networks is also used as model training. The model training process varies depending on the type of problem. In this article, the focus will be on image-based problems, along with the field we call computer vision. Algorithms and architectures consisting of artificial neural networks used in solving image-based problems are referred to as Convolutional Neural Networks in the literature. Convolutional neural networks expect an image as input. The image given to the neural network is processed through layers according to the architecture of that neural network. Since the expected output of the neural network is a group of diseases, it can be said that classification is done for different diseases and that this problem is a classification problem [19].

2.3. SqueezeNet Architecture

Another popular architecture used in the CNN model is the SqueezeNet architecture. The SqueezeNet architecture was introduced by Iandola et al. in 2016. The goal of this architecture is to create a neural network with fewer

parameters, and the architecture achieves AlexNet-level accuracy with 50 times fewer parameters. The advantage of the SqueezeNet architecture is that the workload on the neural network is reduced thanks to more efficient distributed layers, allowing it to run faster.

This architecture was proposed in 2016 by researchers at DeepScale, the University of California, Berkeley, and Stanford University. Its advantages are:

- It requires less communication between servers during distributed training.

- Requires less bandwidth to export a new model from the cloud.

- Is more suitable for deployment on specialized hardware with limited memory. Large activation maps provide higher classification accuracy when given the same number of parameters.

The authors summarize three main strategies for reducing parameter size while maximizing accuracy:

Strategy 1: Make the network smaller by replacing 3x3 filters with 1x1 filters. A 1x1 filter has 9 times fewer parameters than a 3x3 filter.

Strategy 2: Reduce the number of inputs for the remaining 3x3 filters. Fewer inputs to convolutional layers result in fewer parameters.

Strategy 3: Sampling towards the end of the network to ensure the folding layers have large activation maps. This maximizes the use of fewer parameters and achieves the highest accuracy. Delaying subsampling towards the end of the network creates larger activation/feature maps [20].

2.4. Artificial Neural Network (ANN)

The ANN is one of the machine learning methods inspired by biological neural systems. ANNs consist of artificial neurons arranged in layers, and each neuron weights input data and converts it into output through activation functions. This structure allows learning complex nonlinear relationships and achieves high performance in many problems such as classification, regression, and pattern recognition [21].

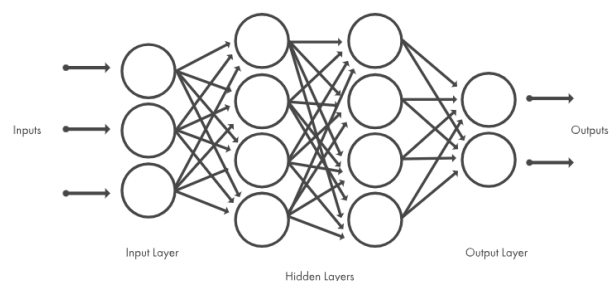


Figure 2. Typical artificial neural network architecture.

2.5. K-Nearest Neighbour (KNN)

The KNN algorithm is a non-parametric, supervised learning classification that makes classifications or predictions about the grouping of individual data points using proximity. It is one of the most popular and simplest

classification and regression algorithms used in machine learning today. While the KNN algorithm can be used in both regression and classification problems, it is most commonly used as a classification algorithm and can assume that similar points are likely to be found close together. Regression problems use a similar concept to the classification problem, but in this case, the average of the k nearest neighbors is taken to make a prediction about a classification. The key difference here is that classification is used for discrete values, while regression is used for continuous values. However, to perform a classification, the distance must first be defined. These distances can be measured in two ways.

Euclidean distance ($p=2$): This is the most commonly used distance measure and is limited to real-valued vectors. Using Formula 1, a straight line is measured between the inquiry point and the other point being measured.

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad (1)$$

Manhattan distance ($p=1$): This is a popular distance measurement that measures the absolute value between two points. Often visualized as a grid, it's also called taxi distance or city block distance, and Formula 2 shows how to get from one address to another via city streets.

$$d(x, y) = \sum_{i=1}^m |x_i - y_i| \quad (2)$$

K-Nearest Neighbor (KNN) applications determine the age (or health) of a new, variable-status tomato plant specimen by calculating the similarity of the feature to labeled specimens shown in your dataset. The algorithm plots the features of the new leaf under examination into the dataspace and identifies "K" neighboring leaves that are most similar to it using persistent distance measurements; it then assigns the new specimen to the class of condition (healthy, early blight, septoria leaf spot, etc.) that is most prevalent among these neighbors. Instead of training a complex weighting model in the background, this method, which directly references probability indicators, physical or pixel-based variation options in your dataset, provides data-driven and direct examples in the diagnosis of tomato diseases.

2.6. Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm frequently used for classification and regression problems in applications such as signal processing, natural language processing (NLP), speech, and image recognition. The goal of the SVM algorithm is to find a hyperplane that best separates data points of one class from data points of another class. This hyperplane can be a line for a 2-dimensional space or a plane for an n -dimensional space, where n is the number of features for each observation in the dataset. There can be more than one hyperplane separating classes in the data. The optimal hyperplane derived by the SVM algorithm is the one that maximizes the margin between the two classes.

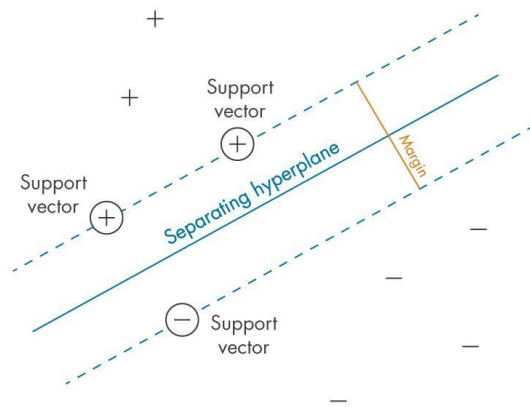


Figure 3. An SVM algorithm to maximize the width of the margin between two classes, represented by plus and minus signs and separated by a hyperplane.

SVM aims to find the ideal decision surface (hyperplane) that defines the boundary between classes by plotting the features obtained from leaves (color change, texture distortions, spot size, etc.) in a multidimensional space. Especially in cases where disease symptoms are complex and cannot be linearly separated by a simple line, it uses "kernel functions" to move the data to higher dimensions, thus mathematically drawing the most precise distinction between early blight, leaf spot, or healthy states; this allows the algorithm to achieve high accuracy rates and strong generalization capabilities in the early diagnosis of diseases when a new tomato leaf image is presented to it.

2.7. Logistic Regression (LR)

Logistic regression learns by providing the system with the necessary class information corresponding to the input values related to the problem. The model obtained as a result of this process can make meaningful predictions for any input value that it has not encountered before and for which it does not have class information, thanks to its generalization ability. When modeling, it aims to find the most suitable function in order to reach the closest possible result. Because the error returns to the system as learning, the process is much faster than unsupervised learning [22]. Logistic regression is a method that has the logic of creating a relationship between the dependent variable and the independent variable over a small number of variables. The most important feature that distinguishes logistic regression from linear regression is that modeling can be achieved without the need for criteria such as the normal distribution, constant variance, and linearity of the variable [22].

2.8. Confusion Matrix

A confusion matrix is a visualization method for classifier algorithm results. More specifically, it is a table that divides the number of actual instances of a given class by the number of predicted instances of that class. Confusion matrices are one of several evaluation metrics

that measure the performance of a classification model. They can be used to calculate many other model performance metrics such as accuracy and recall. Confusion matrices can be used with any classifier algorithm, such as logistic regression models, decision trees, etc.

In a confusion matrix, the columns represent the predicted values of a given class, and the rows represent the actual values of a given class, or vice versa. This grid structure is a useful tool for visualizing model classification accuracy by showing the number of correct and incorrect predictions side-by-side for all classes.

		PREDICTED	
ACTUAL		Tomato - Bacterial Spot	Tomato - Healthy
	Tomato - Bacterial Spot	TP	FP
	Tomato - Healthy	FN	TN

Figure 5. An example of a Confusion Matrix.

2.9. Performance Metrics

2.9.1. Accuracy

It shows the percentage of all predictions made by the model that are correct. It is the simplest and most common metric.

True Positive (TP) is the case where samples that are actually positive are correctly predicted as positive. True Negative (TN) is the case where samples that are actually negative are correctly predicted as negative. False Positive (FP) is the case where samples that are actually negative are incorrectly predicted as positive. False Negative (FN) indicates the case where samples that are actually positive are incorrectly predicted as negative.

The accuracy metric evaluates the probability of the model making the correct prediction by considering these four cases. However, accuracy alone may be insufficient in situations such as unbalanced datasets or cost-focused problems. Therefore, it is more appropriate to use it in conjunction with other metrics to evaluate the model's performance in more detail.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

All models demonstrated exceptional overall performance, achieving an accuracy rate above 0.99. In particular, the Artificial Neural Network (ANN) model achieved an accuracy of 1.000 (close to 100%), proving its near-perfect prediction of the leaf condition in the tested dataset. These high accuracy rates clearly demonstrate that the developed system can be used as a reliable early

detection tool in greenhouses or fields.

2.9.2. F1 Score

The F1 score is the harmonic mean of the Precision and Recall metrics. It provides a better measure than Accuracy, especially in unbalanced datasets, because it takes both metrics into account in a balanced way. This is shown in Formula 4.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The F1 score is proof that the model does not overweight one metric while neglecting the other. The near-perfect F1 scores of 0.996 for ANN and 0.994 for SVM confirm that the models successfully detect diseased leaves with over 99% accuracy and avoid false alarms for healthy leaves. The ANN and SVM models provided the most optimal balance between the two main objectives of the study: "early detection" and "chemical minimization".

2.9.3. Precision

Sensitivity indicates how many of the cases that the model labels as positive are actually positive. This is quite important in situations where the cost of false positives (FP) is high. It is shown in Formula 5.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

If a model has low sensitivity, it can cause unnecessary chemical spraying by labeling healthy plants as "sick" (False Positive). Examining the table, the lowest sensitivity is found in the KNN algorithm with 0.952; meaning that approximately 4.8% of the "diseased" alarms given by the KNN model are actually for healthy leaves. This can lead to unnecessary spraying. In contrast, the ANN (0.986) and SVM (0.979) models minimize the "false alarm" rate with very high sensitivity values. In this way, chemical pesticide use can be significantly reduced by targeting only truly diseased tomato plants, thus achieving the goal of environmentally friendly agriculture (sustainability).

2.9.4. Recall

It shows the percentage of actually positive cases that were correctly identified by the model. It can be used in situations where the cost of false negatives (FN) is high. This is shown in Formula 6.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

The greatest danger in agricultural diseases is that a diseased plant is overlooked and mistaken for a "healthy" plant (False Negative), leading to the spread of the disease throughout the greenhouse. Looking at the sensitivity values of the models, ANN (0.993) yielded the most successful result. Interestingly, the KNN model, which is weakest in precision, is quite successful in sensitivity (0.990); that is, it works well in not missing diseases, but in doing so it can be a little too "cautious" and mistake healthy plants for diseased

ones. These high recall values obtained to prevent the spread of disease demonstrate that the system has a very strong early diagnosis capability.

3. Results And Discussion

This study aimed to classify different tomato plants by detecting disease symptoms using image processing and machine learning methods, employing ANN, KNN, SVM, and LR techniques with the Crop Disease Detection dataset. A computer with a Core i5-7200U 2.50 GHz processor, Intel(R) HD Graphics 620 graphics card, and 8 GB of RAM was used. The Python programming language was employed. For the ANN method, the following parameters were used: 100 hidden layers, ReLu activation function, regularization 0.0001, and number of iterations 200. For the KNN method, the following parameters were used: number of neighbors 5, metric euclidean, and weight uniform. For the SVM method, the following parameters were used: cost 1, regression loss epsilon 0.1, kernel RBF, and iteration 100. For the LR method, the regularization type Ridge L2 parameter was used. Training and testing were performed using cross-validation, with a k value of 10. The processes performed in this study are shown in Figure 6.

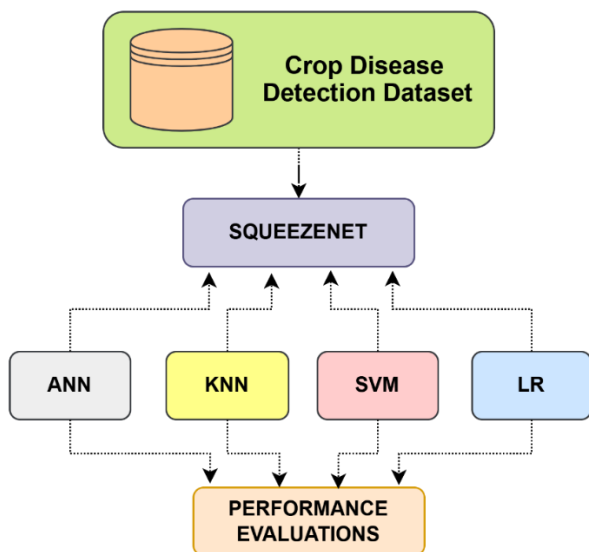


Figure 6. Flowchart of the steps performed in the study.

As an introduction to machine learning methods, the SqueezeNet model was used to extract approximately 6 features from each image, and training was performed. The confusion matrices obtained from the machine learning methods are given below. The confusion matrix obtained from the ANN method is

shown in Figure 7.

		PREDICTED					
ACTUAL		Tomato - Bacterial Spot	Tomato - Early Blight	Tomato - Healthy	Tomato - Late Blight	Tomato - Septoria Leaf Spot	Tomato - Yellow Leaf Curl Virus
	Tomato - Bacterial Spot	2068	18	0	4	18	19
	Tomato - Early Blight	31	2229	15	60	46	19
	Tomato - Healthy	0	6	2390	4	7	0
	Tomato - Late Blight	4	68	11	2199	26	6
	Tomato - Septoria Leaf Spot	18	50	8	19	2689	0
	Tomato - Yellow Leaf Curl Virus	19	20	1	4	1	2406

Figure 7. Confusion Matrix of the ANN Model

According to Figure 7, the ANN method correctly classified 2068 data points belonging to the Tomato-Bacterial Spot category. It classified 18 data points belonging to the Tomato-Bacterial Spot category as Tomato-Early Blight. It classified 4 data points belonging to the Tomato-Bacterial Spot category as Tomato-Late Blight. It classified 18 data points belonging to the Tomato-Bacterial Spot category as Tomato-Septoria Leaf Spot. It classified 19 data points belonging to the Tomato-Bacterial Spot category as Tomato-Yellow Leaf Curl Virus. In total, it incorrectly classified 59 data points belonging to the Tomato-Bacterial Spot category.

The study correctly classified 2229 data points belonging to the "Tomato - Early Blight" category. It classified 31 data points belonging to the "Tomato - Bacterial Spot" category. It classified 15 data points belonging to the "Tomato - Early Blight" category as "Tomato - Healthy". It classified 60 data points belonging to the "Tomato - Early Blight" category as "Tomato - Late Blight". It classified 46 data points belonging to the "Tomato - Early Blight" category as "Tomato - Septoria Leaf Spot". It classified 19 data points belonging to the "Tomato - Early Blight" category as "Tomato - Yellow Leaf Curl Virus". In total, it incorrectly classified 171 data points belonging to the "Tomato - Early Blight" category.

The study correctly classified 2390 data points belonging to the "Healthy Tomato" category. It classified 6 data points belonging to the "Healthy Tomato" category as "Early Blight," 4 data points as "Late Blight," and 7 data points as "Septoria Leaf Spot." In total, it incorrectly classified 17 data points belonging to the "Healthy Tomato" category.

The study correctly classified 2199 data points belonging to the "Tomato - Late Blight" category. It classified 4 data points belonging to the "Tomato - Late Blight" category as "Tomato - Bacterial Spot". It classified 68 data points belonging to the "Tomato - Late Blight" category as "Tomato - Early Blight". It classified 11 data points belonging to the "Tomato - Late Blight" category as "Tomato - Healthy". It classified 26 data points belonging to the "Tomato - Late Blight" category as "Tomato -

Septoria Leaf Spot". It classified 6 data points belonging to the "Tomato - Late Blight" category as "Tomato - Yellow Leaf Curl Virus". In total, it incorrectly classified 115 data points belonging to the "Tomato - Early Blight" category.

The study correctly classified 2089 data points belonging to the Tomato-Septoria Leaf Spot category. It classified 18 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Bacterial Spot. It classified 50 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Early Blight. It classified 8 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Healthy. It classified 19 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Late Blight. In total, it incorrectly classified 95 data points belonging to the Tomato-Septoria Leaf Spot category.

The study correctly classified 2406 data points belonging to the Tomato-Yellow Leaf Curl Virus class. It classified 19 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Bacterial Spot. It classified 20 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Early Blight. It classified 1 data point belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Healthy. It classified 4 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Late Blight. It classified 1 data point belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Septoria Leaf Spot. In total, it incorrectly classified 45 data points belonging to the Tomato-Yellow Leaf Curl Virus class.

	PREDICTED					
	Tomato - Bacterial Spot	Tomato - Early Blight	Tomato - Healthy	Tomato - Late Blight	Tomato - Septoria Leaf Spot	Tomato - Yellow Leaf Curl Virus
Tomato - Bacterial Spot	2052	15	8	11	11	30
Tomato - Early Blight	323	1829	57	68	100	23
Tomato - Healthy	5	12	2382	0	7	1
Tomato - Late Blight	125	159	63	1876	59	32
Tomato - Septoria Leaf Spot	221	104	76	60	1705	18
Tomato - Yellow Leaf Curl Virus	98	21	12	9	9	2302

Figure 8. Confusion Matrix of the KNN Model

According to Figure 8, the KNN method correctly classified 2052 data points belonging to the Tomato-Bacterial Spot category. It classified 15 data points as Tomato-Early Blight, 8 as Tomato-Healthy, 11 as Tomato-Late Blight, 11 as Tomato-Septoria Leaf Spot, and 30 as Tomato-Yellow Leaf Curl Virus. In total, it incorrectly classified 75 data points belonging to the Tomato-Bacterial Spot category.

The study correctly classified 1829 data points belonging to the "Tomato - Early Blight" category. It classified 323 data points belonging to the "Tomato - Bacterial Spot" category. It classified 57 data points

belonging to the "Tomato - Early Blight" category as "Tomato - Healthy". It classified 68 data points belonging to the "Tomato - Early Blight" category as "Tomato - Late Blight". It classified 100 data points belonging to the "Tomato - Early Blight" category as "Tomato - Septoria Leaf Spot". It classified 23 data points belonging to the "Tomato - Early Blight" category as "Tomato - Yellow Leaf Curl Virus". In total, it incorrectly classified 571 data points belonging to the "Tomato - Early Blight" category.

The study correctly classified 2382 data points belonging to the "Healthy Tomato" category. It classified 5 data points from the "Healthy Tomato" category as "Bacterial Spot". It classified 12 data points from the "Healthy Tomato" category as "Early Blight". It classified 7 data points from the "Healthy Tomato" category as "Septoria Leaf Spot". It classified 1 data point from the "Healthy Tomato" category as "Yellow Leaf Curl Virus". In total, it incorrectly classified 25 data points from the "Healthy Tomato" category.

The study correctly classified 1876 data points belonging to the "Tomato - Late Blight" category. It classified 125 data points belonging to the "Tomato - Bacterial Spot" category. It classified 159 data points belonging to the "Tomato - Late Blight" category as "Tomato - Early Blight". It classified 63 data points belonging to the "Tomato - Healthy" category. It classified 59 data points belonging to the "Tomato - Late Blight" category as "Tomato - Septoria Leaf Spot". It classified 32 data points belonging to the "Tomato - Late Blight" category as "Tomato - Yellow Leaf Curl Virus". In total, it incorrectly classified 438 data points belonging to the "Tomato - Early Blight" category.

The study correctly classified 1705 data points belonging to the Tomato-Septoria Leaf Spot category. It classified 221 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Bacterial Spot. It classified 104 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Early Blight. It classified 76 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Healthy. It classified 60 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Late Blight. It classified 18 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Yellow Leaf Curl Virus. In total, it incorrectly classified 479 data points belonging to the Tomato-Septoria Leaf Spot category.

The study correctly classified 2302 data points belonging to the Tomato-Yellow Leaf Curl Virus class. It classified 98 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Bacterial Spot. It classified 21 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Early Blight. It classified 12 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Healthy. It classified 9 data points belonging to the Tomato-Yellow Leaf Curl Virus class as

Tomato-Late Blight. It classified 9 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Septoria Leaf Spot. In total, it incorrectly classified 149 data points belonging to the Tomato-Yellow Leaf Curl Virus class.

		PREDICTED					
ACTUAL		Tomato - Bacterial Spot	Tomato - Early Blight	Tomato - Healthy	Tomato - Late Blight	Tomato - Septeria Leaf Spot	Tomato - Yellow Leaf Curl Virus
	Tomato - Bacterial Spot	1979	44	0	7	66	31
	Tomato - Early Blight	90	1958	24	175	128	25
	Tomato - Healthy	0	8	2379	5	15	0
	Tomato - Late Blight	40	286	13	1876	95	4
	Tomato - Septeria Leaf Spot	77	189	14	42	1659	3
	Tomato - Yellow Leaf Curl Virus	34	32	0	5	2	2378

Figure 9. Confusion Matrix of the SVM Model

According to Figure 9, the SVM method correctly classified 1979 data points belonging to the Tomato-Bacterial Spot category. It classified 44 data points belonging to the Tomato-Bacterial Spot category as Tomato-Early Blight. It classified 7 data points belonging to the Tomato-Bacterial Spot category as Tomato-Late Blight. It classified 66 data points belonging to the Tomato-Bacterial Spot category as Tomato-Septoria Leaf Spot. It classified 31 data points belonging to the Tomato-Bacterial Spot category as Tomato-Yellow Leaf Curl Virus. In total, it incorrectly classified 148 data points belonging to the Tomato-Bacterial Spot category.

The study correctly classified 1958 data points belonging to the "Tomato - Early Blight" category. It classified 90 data points belonging to the "Tomato - Bacterial Spot" category. It classified 24 data points belonging to the "Tomato - Early Blight" category as "Tomato - Healthy". It classified 175 data points belonging to the "Tomato - Early Blight" category as "Tomato - Late Blight". It classified 128 data points belonging to the "Tomato - Early Blight" category as "Tomato - Septoria Leaf Spot". It classified 25 data points belonging to the "Tomato - Early Blight" category as "Tomato - Yellow Leaf Curl Virus". In total, it incorrectly classified 442 data points belonging to the "Tomato - Early Blight" category.

The study correctly classified 2379 data points belonging to the "Healthy Tomato" category. It classified 8 data points belonging to the "Healthy Tomato" category as "Early Blight," 5 data points as "Late Blight," and 15 data points as "Septoria Leaf Spot." In total, it incorrectly classified 28 data points belonging to the "Healthy Tomato" category.

The study correctly classified 1876 data points belonging to the "Tomato - Late Blight" category. It classified 40 data points belonging to the "Tomato - Bacterial Spot" category. It classified 286 data points

belonging to the "Tomato - Late Blight" category as "Tomato - Early Blight". It classified 13 data points belonging to the "Tomato - Healthy" category. It classified 95 data points belonging to the "Tomato - Late Blight" category as "Tomato - Septoria Leaf Spot". It classified 4 data points belonging to the "Tomato - Late Blight" category as "Tomato - Yellow Leaf Curl Virus". In total, it incorrectly classified 438 data points belonging to the "Tomato - Early Blight" category.

The study correctly classified 1859 data points belonging to the Tomato-Septoria Leaf Spot category. It classified 77 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Bacterial Spot. It classified 189 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Early Blight. It classified 14 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Healthy. It classified 42 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Late Blight. It classified 3 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Yellow Leaf Curl Virus. In total, it incorrectly classified 325 data points belonging to the Tomato-Septoria Leaf Spot category.

The study correctly classified 2378 data points belonging to the Tomato-Yellow Leaf Curl Virus class. It classified 34 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Bacterial Spot. It classified 32 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Early Blight. It classified 5 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Late Blight. It classified 2 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Septoria Leaf Spot. In total, it incorrectly classified 73 data points belonging to the Tomato-Yellow Leaf Curl Virus class.

		PREDICTED					
ACTUAL		Tomato - Bacterial Spot	Tomato - Early Blight	Tomato - Healthy	Tomato - Late Blight	Tomato - Septeria Leaf Spot	Tomato - Yellow Leaf Curl Virus
	Tomato - Bacterial Spot	2017	35	1	16	28	30
	Tomato - Early Blight	55	2103	17	142	64	19
	Tomato - Healthy	0	16	2352	20	19	0
	Tomato - Late Blight	14	144	17	2070	52	17
	Tomato - Septeria Leaf Spot	27	78	16	44	2017	2
	Tomato - Yellow Leaf Curl Virus	30	22	1	20	3	2375

Figure 10. Confusion Matrix for the LR model.

According to Figure 10, the LR method correctly classified 2017 data points belonging to the Tomato-Bacterial Spot category. It classified 35 data points as Tomato-Early Blight. It classified 1 data point as Tomato-Healthy. It classified 16 data points as Tomato-Late Blight. It classified 28 data points as Tomato-Septoria Leaf Spot. It classified 30 data points as Tomato-Yellow Leaf Curl

Virus. In total, it incorrectly classified 110 data points belonging to the Tomato-Bacterial Spot category.

The study correctly classified 2103 data points belonging to the "Tomato - Early Blight" category. It classified 55 data points belonging to the "Tomato - Bacterial Spot" category. It classified 17 data points belonging to the "Tomato - Early Blight" category as "Tomato - Healthy". It classified 142 data points belonging to the "Tomato - Early Blight" category as "Tomato - Late Blight". It classified 64 data points belonging to the "Tomato - Early Blight" category as "Tomato - Septoria Leaf Spot". It classified 19 data points belonging to the "Tomato - Early Blight" category as "Tomato - Yellow Leaf Curl Virus". In total, it incorrectly classified 155 data points belonging to the "Tomato - Early Blight" category.

The study correctly classified 2352 data points belonging to the "Healthy Tomato" category. It classified 16 data points belonging to the "Healthy Tomato" category as "Early Blight," 20 data points as "Late Blight," and 19 data points as "Septoria Leaf Spot." In total, it incorrectly classified 55 data points belonging to the "Healthy Tomato" category.

The study correctly classified 2070 data points belonging to the "Tomato - Late Blight" category. It classified 14 data points belonging to the "Tomato - Bacterial Spot" category. It classified 144 data points belonging to the "Tomato - Late Blight" category as "Tomato - Early Blight". It classified 17 data points belonging to the "Tomato - Healthy" category. It classified 52 data points belonging to the "Tomato - Late Blight" category as "Tomato - Septoria Leaf Spot". It classified 17 data points belonging to the "Tomato - Late Blight" category as "Tomato - Yellow Leaf Curl Virus". In total, it incorrectly classified 244 data points belonging to the "Tomato - Early Blight" category.

The study correctly classified 2017 data points belonging to the Tomato-Septoria Leaf Spot category. It classified 27 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Bacterial Spot. It classified 78 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Early Blight. It classified 16 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Healthy. It classified 44 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Late Blight. It classified 2 data points belonging to the Tomato-Septoria Leaf Spot category as Tomato-Yellow Leaf Curl Virus. In total, it incorrectly classified 167 data points belonging to the Tomato-Septoria Leaf Spot category.

The study correctly classified 2375 data points belonging to the Tomato-Yellow Leaf Curl Virus class. It classified 30 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Bacterial Spot. It classified 22 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Early Blight. It classified

1 data point belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Healthy. It classified 20 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Late Blight. It classified 3 data points belonging to the Tomato-Yellow Leaf Curl Virus class as Tomato-Septoria Leaf Spot. In total, it incorrectly classified 76 data points belonging to the Tomato-Yellow Leaf Curl Virus class.

Performance metrics for each method were calculated using data from complex matrices obtained from the methods. These metrics are shown in Table 1.

Table 1. Performance Metrics for All Models

	Accuracy	F1 Score	Precision	Recall
KNN	0,998	0,983	0,952	0,990
LR	0,999	0,992	0,978	0,977
SVM	0,999	0,994	0,979	0,988
ANN	1,000	0,996	0,986	0,993

According to Table 1, the most successful model is ANN with a score of 1.000. The model with the lowest classification success is the KNN model with a score of 0.998. The classification success of the LR model is 0.999, and the classification success of the SVM model is also 0.999. When other performance metrics are examined, it is seen that the F1 Score, Precision, and Recall metric values also show parallelism with the classification success of the models. When the classification successes and F1 Score, precision, and recall metrics are examined, it can be said that all models show a success rate of 99.8% and above. Figure 11 shows the ROC analysis curve for all models of the tomato image data titled "Tomato - Bacterial Spot".

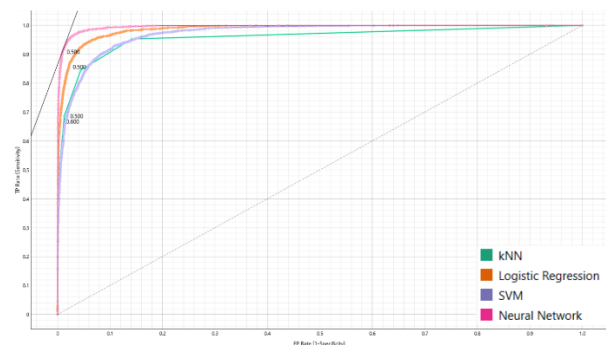


Figure 11. Tomato - ROC analysis curve of early blight image data.

Figure 11 shows a parallel between the classification success rates and the learning levels of the models in the Tomato Early Blight data. ANN is the best-learning model, while KNN is the least-learning model. Therefore, it can be said that the classification success of the models is determined by their learning rates. After classifying the

data, the main goal is to improve the health of tomato plants by reducing the amount of pesticides applied and implementing organic farming methods on tomato plants.

After the classification of tomato plant data is completed, certain tests must be performed and care must be taken in cultivating these plants to prevent these diseases from occurring. As a result, the bacteria that cause the first disease, Bacterial Spot, live on plant debris. The onset of the disease is facilitated by hot and rainy weather. The symptoms of Bacterial Spot are very similar to other tomato diseases, but only bacterial spots cause bacterial ooze from a cut leaf. The disease spreads through infected seeds, wind-driven rain, diseased seedlings, or infested soil. Bacteria can enter the plant through natural openings in the leaves or through openings caused by damage to the leaves. To prevent Bacterial Spot disease, it is necessary to use only certified seeds and healthy seedlings and to clean all crop residues from the planting area. Additionally, watering the plant at the base instead of using sprinklers and practicing crop rotation are measures taken to prevent tomato plants from contracting Bacterial Spot disease [18].

Another disease, Early Blight, can spread rapidly after plants bear fruit; airborne spore movement and contact with infected soil can cause the disease to spread. Applying appropriate fungicides at the first sign of the disease, eradicating certain wild-growing nightshade plants (tomatoes, potatoes, etc.), and implementing crop rotation can be effective in preventing this disease.

Late Blight disease, which is a distinct disease, can destroy all tomato cultivation areas where it occurs. Therefore, resistant tomato varieties should be planted to prevent Late Blight disease. If disease symptoms are present, or if rainy conditions are likely or overhead irrigation is used, appropriate fungicides should be applied in specific quantities and in a knowledgeable manner [18].

Septoria Leaf Spot disease can spread through water splashing. The fungus that causes this disease can survive the winter in plant debris. To prevent this disease, ensure that all tomato crop debris is cleaned up and destroyed or thoroughly mixed into the soil in the fall. Planting only disease-free plants, avoiding overhead irrigation, and staking plants to increase air circulation between leaves are measures taken to prevent the disease. As a last resort, applying the appropriate amount of fungicide prevents the disease.

The Yellow Leaf Curl Virus (YLCV) is transmitted by whiteflies and can cause a 100% yield loss if it infects plants during the early stages of the crop. The virus can also infect other hosts such as beans, ornamental plants, and various weed species. To prevent this disease, resistant tomato varieties can be grown. It is necessary to plant disease- and whitefly-free seedlings, remove and burn infected plants, and clear the field of weeds. Measures such as using yellow sticky traps to monitor and control whiteflies can be taken. If the insect infestation is severe,

spraying appropriate insecticides is one of the measures taken to prevent the disease.

It is crucial to practice organic farming consciously to limit the use of pesticides and avoid overuse. This study sought to determine what kind of disease the tomato plant data used had on tomatoes. To this end, different tomato plant data were examined and classified using specific methods. The method that applied the classification most accurately was ANN with 100%. The lowest classification success was KNN with 99.8%.

After identifying the disease with these data, the study explains what can be done to prevent it. The aim is to reduce pesticide use or limit it within certain boundaries. However, with more data, all tomato diseases can be identified, and thus pesticide use can be gradually reduced, promoting the spread of organic farming and ensuring the health of living beings and a cleaner environment.

The classifications recommended in the study are used in greenhouses and gardens where tomato plants are grown, facilitating the identification of tomato plant diseases and helping to take appropriate measures. This enables the cultivation of healthier tomato plants by preventing the indiscriminate use of pesticides.

4. Conclusions

It is crucial to practice organic farming consciously to limit the use of pesticides and avoid overuse. This study sought to determine what kind of disease the tomato plant data used had on tomatoes. To this end, different tomato plant data were examined and classified using specific methods. The method that applied the classification most accurately was ANN with 100%. The lowest classification success was KNN with 99.8%.

After identifying the disease with these data, the study explains what can be done to prevent it. The aim is to reduce pesticide use or limit it within certain boundaries. However, with more data, all tomato diseases can be identified, and thus pesticide use can be gradually reduced, promoting the spread of organic farming and ensuring the health of living beings and a cleaner environment.

The classifications recommended in the study are used in greenhouses and gardens where tomato plants are grown, facilitating the identification of tomato plant diseases and helping to take appropriate measures. This enables the cultivation of healthier tomato plants by preventing the indiscriminate use of pesticides.

Declaration of Ethical Standards

The article does not contain any studies with human or animal subjects.

Credit Authorship Contribution Statement

Authors individually were responsible for the ideation, modeling, analysis, and writing of this article.

Declaration of Competing Interest

Author claims that there are no conflicts of interest.

Funding / Acknowledgements

No funding was received from any organization for this research.

Data Availability

<https://www.kaggle.com/datasets/snikhilrao/crop-disease-detection-dataset>

References

- [1] A. Akdoğan, Ü. Divrikli, and L. Elçi, "Importance of pesticides and their effects on ecosystem," *Akademik Gıda*, vol. 10, no. 1, pp. 125–132, Mar. 2012. [Online]. Available: <https://izlik.org/JA26ND97UJ>
- [2] O. Tiryaki, R. Canhalil, and S. Horuz, "The use of pesticides and their risks," *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi*, vol. 26, no. 2, pp. 154–169, Apr. 2010. [Online]. Available: <https://izlik.org/JA67XU23RC>
- [3] Z. Özcan and S. Tongur, "Pestisitlerin toksisitesinin *Lepidium sativum* test moduyla çevre ve insan sağlığı açısından değerlendirilmesi," *Ulusal Çevre Bilimleri Araştırma Dergisi*, vol. 2, no. 4, pp. 144–150, 2019.
- [4] Ş. E. Arıcı et al., "İsparta/Keçiborlu ilçesinde yağ gülü yetiştiriciliği, pestisit kullanımı hakkında üretici sorunları ve uygulama esasları," *Turkish Journal of Science and Engineering*, vol. 4, no. 1, pp. 15–20, 2022.
- [5] R. Yüzbaşıoğlu and Ş. Topkaya, "Üreticilerin tarım ilacı uygulamasındaki bilgisi ve ilaçlamada çevre duyarlılığı: Tokat ili merkez ilçe örneği," *Gaziosmanpaşa Bilimsel Araştırma Dergisi*, vol. 11, no. 3, pp. 306–316, 2022.
- [6] Ö. Ü. A. Polat, "Tarımda pestisitler: Dünya'da ve Türkiye'de kullanımları," 2022.
- [7] B. Özercan and R. Taşcı, "Türkiye'de pestisit kullanımının iller, bölgeler ve pestisit grupları açısından incelenmesi," *Ziraat Mühendisliği*, no. 375, pp. 75–88, 2022.
- [8] H. Yılmaz, A. Düzenli, and M. M. Dağ, "Dünya, Avrupa Birliği ülkeleri ve Türkiye'de pestisit kullanımı ve yasal düzenlemeler," *Türkiye Tarımsal Araştırmalar Dergisi*, vol. 11, no. 3, pp. 315–330, 2024.
- [9] M. Ayyıldız, "Türkiye'de kimyasal pestisit kullanımının ekonomi ve çevre yönüyle değerlendirmesi," *COMU Journal of Agriculture Faculty*, vol. 10, no. 2, pp. 267–274, 2022.
- [10] H. H. Tıraş, "Tarım üretiminde önde gelen ülkelerde tarımsal gübre ve ilaç kullanımının tarımsal GSYİH'ye etkisi: Panel veri analizi," *Business & Management Studies: An International Journal*, vol. 12, no. 1, pp. 132–147, 2024.
- [11] F. N. Doğan and M. E. Karpuzcu, "En iyi yönetim uygulamalarının pestisit kirliliğinin kontrolünde kullanımı," *Çevre İklim ve Sürdürülebilirlik*, vol. 24, no. 1, pp. 25–32, 2023.
- [12] M. Aydoğan and M. Baran, "Tütün yetiştiriciliğinde tarım ilaçları kullanımı ve üreticilerin zirai mücadele bilinç düzeylerinin belirlenmesi: Çelikhan tütün örneği, Türkiye," *Türkiye Tarımsal Araştırmalar Dergisi*, vol. 10, no. 2, pp. 180–190, 2023.
- [13] S. Hayran, A. D. Çelik, and A. Gül, "Türkiye'de sera çiftçilerinin pestisit kullanımı ve risk algılamaları: Mersin ili örneği," *Tarım Ekonomisi Dergisi*, vol. 28, no. 1, pp. 31–39, 2022.
- [14] H. G. Doğan and H. Karaaslan, "Türkiye'de tarım sektörü kaynaklı çevre tahribatını önleme politikaları kapsamında destekleme alan üreticilerin incelenmesi (Seyfe Gölü yöresi hububat üreticileri örneği)," *Turkish Journal of Agriculture-Food Science and Technology*, vol. 10, no. 11, pp. 2213–2223, 2022.
- [15] M. Tözün and G. Akar, "Türkiye'de gıda numunelerinde pestisit kalıntıları üzerine 2010 yılı sonrası ulusal literatürün incelenmesi," *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Halk Sağlığı Dergisi*, vol. 7, no. 1, pp. 177–191, 2022.
- [16] O. Gedikli, A. Uzundumlu, and G. Tozlu, "Çeltik, mısır ve buğday üretiminde tarımsal ilaç kullanımının çevresel duyarlılık yönünden incelenmesi: Samsun ili örneği," *TÜBAV Bilim Dergisi*, vol. 8, no. 2, pp. 19–26, 2015.
- [17] D. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," *arXiv preprint arXiv:1511.08060*, 2015.
- [18] F. A. Esen and A. Onan, "Derin öğrenme yöntemleri ile bitki yaprakları üzerindeki hastalıkların sınıflandırılması," *Avrupa Bilim ve Teknoloji Dergisi*, no. 40, pp. 151–155, 2022.
- [19] S. Gökarp and İ. Aydın, "Farklı derin sinir ağı modellerinin duygu tanımadaki performanslarının karşılaştırılması," *Muş Alparslan Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, vol. 2, no. 1, pp. 35–43, 2021.
- [20] Y. S. Taspinar, "Machine learning-driven e-nose-based diabetes detection: Sensor selection and feature reduction study," *Sensors*, vol. 25, no. 21, p. 6607, 2025.
- [21] B. Bahçıvan and A. Yılmaz, "Demografik özelliklerin çevrimiçi market alışverişi kullanımına etkisinin makine öğrenmesi yöntemleri ile tahmini," *Konya Journal of Engineering Sciences*, vol. 11, no. 4, pp. 942–957, 2023.

*Research Article***Simulation-Based Evaluation of LiDAR–Photogrammetry Fusion via NeRF Reconstruction and ICP Registration in Urban Scenes**

Rytis Maskeliūnas ^{a,*} , **Sarmad Maqsood** ^{a,} , **Ahmad Qurthobi** ^{a,} , **Irfan Abbas** ^{a,} ,
Mantas Vaškevičius ^{b,} , **Julius Gelšvartas** ^{b,}

^a Centre of Real Time Computer Systems, Faculty of Informatics, Kaunas University of Technology, LT-51386 Kaunas, Lithuania.

^b Matomai UAB, LT-51423 Kaunas, Lithuania

ARTICLE INFO

Article history:

Received 28 January 2026

Accepted 9 March 2026

Keywords:

NeRF,
LiDAR–Photogrammetry
Fusion,
Point-Cloud Registration,
Simulation,
Urban Scenes

ABSTRACT

Reliable integration of LiDAR and photogrammetric point clouds is essential for high-precision 3D mapping, yet systematic evaluations of fusion accuracy under controlled conditions remain limited. This study presents a simulation-based assessment framework for LiDAR–photogrammetry fusion using a neural radiance field (NeRF) representation and iterative closest point (ICP) registration. Experiments were conducted in the CARLA simulator (Town 10), where a drone-mounted multi-sensor platform (RGB, LiDAR, GPS, IMU) generated spatially aligned datasets. Photogrammetric reconstructions were produced using the Nerfstudio–Nerfacto pipeline with varied architectural and sampling configurations to analyze their impact on cross-modal registration. Quantitative evaluation employed Chamfer and cloud-to-cloud (C2C) distances to assess NeRF reconstruction fidelity, and ICP root-mean-square error (RMSE), inlier fitness, and runtime to evaluate registration performance. Results show that tuning NeRF’s hidden dimensions and sampling levels yields up to 25% lower ICP RMSE and faster convergence across object categories. The proposed framework enables reproducible benchmarking of LiDAR–photogrammetry fusion and provides a foundation for extending NeRF-based methods to real-world urban mapping scenarios.

This is an open access article under the CC BY-SA 4.0 license.
(<https://creativecommons.org/licenses/by-sa/4.0/>)

1. Introduction

High-fidelity 3D mapping and reconstruction are critical in robotics, geospatial sensing, autonomous navigation, and smart cities [1], [2]. Light detection and ranging (LiDAR) sensors offer precise depth measurements and geometric accuracy even in low texture environments but typically lack rich color or texture information [3]. Photogrammetry, on the other hand, reconstructs dense, textured surfaces from image collections but often suffers from scale drift, holes in textureless or repetitive-surface regions, and sensitivity to lighting and viewpoint. A promising approach is to fuse LiDAR and photogrammetric point clouds to combine geometric accuracy with dense appearance and surface

detail. The idea of hybrid LiDAR-photogrammetry fusion has been explored in prior works [4].

However, achieving robust and accurate fusion remains nontrivial. Disparate sensor modalities lead to differences in point densities, noise characteristics, and coordinate referencing [5]. Photogrammetry reconstructions may drift spatially or warp subtly, particularly in textureless or weakly constrained regions. Likewise, LiDAR scans carry measurement noise and partial occlusions [6], [7], [8]. Standard registration methods, such as iterative closest point (ICP), are widely used but can fail under poor initialization, noise, or low overlap; in particular, ICP is known to be sensitive to local minima and initialization quality [9].

Recently, neural radiance field (NeRF) representations

* Corresponding author. E-mail address: rytis.maskeliunas@ktu.lt
DOI: 10.58190/ijamec.2026.162

have emerged as a powerful alternative to traditional photogrammetric reconstruction, enabling continuous, photorealistic scene modeling and implicit geometry encoding. Approaches such as bundle-adjusting (BARF) NeRF [10] jointly optimize camera poses and scene radiance fields, improving geometric consistency, while DReg-NeRF [11] extends this idea toward learned registration between neural fields. Nevertheless, prior literature rarely isolates where the true bottleneck lies, whether in the reconstruction quality, the registration algorithm, or their interaction across sensing modalities. This lack of controlled evaluation motivates our work.

In this work, we address that gap by providing a controlled simulation-based evaluation of LiDAR–photogrammetry fusion. Our contributions are:

A reproducible CARLA-based testbed with a drone-mounted sensor suite (RGB, LiDAR, GPS, IMU) in urban scenes, enabling tight control over trajectories and noise.

An ablation study of NeRF-based photogrammetric reconstruction (varying hidden dimension, sampling levels, etc.) to quantify how reconstruction quality affects alignment.

Quantitative fusion benchmarks, reporting Chamfer / cloud-to-cloud (C2C) distances for reconstruction fidelity, and ICP RMSE, inlier fitness, and runtime for registration, including a noise-sensitivity analysis.

Our results show that properly tuned NeRF reconstructions can reduce ICP registration error by up to ~25% compared to naive or default configurations, yield more stable convergence across object types (e.g. poles, road signs, benches), and maintain viable runtimes, underscoring the importance of reconstruction design in effective LiDAR–photogrammetry fusion.

2. Related works

2.1. Photogrammetry and NeRF for 3D Reconstruction

Traditional photogrammetry frameworks reconstruct dense surface models from overlapping 2-D images using structure-from-motion and multi-view-stereo algorithms [12]. Despite high visual fidelity, these methods often suffer from scale drift, poor depth recovery in textureless regions, and lighting sensitivity. Recent advances in neural implicit representations, particularly the NeRF [13], overcome many of these limitations by encoding scene geometry and appearance within a continuous volumetric function. Follow-up variants such as BARF NeRF [10] and DReg-NeRF [11] introduced joint camera-pose optimization and learned inter-NeRF registration, respectively, further improving geometric consistency. While these methods excel at reconstructing compact scenes, few studies have quantitatively examined their alignment behavior when fused with external LiDAR data under controlled noise conditions.

2.2. LiDAR–Image Fusion and Registration

Integrating LiDAR with optical data has long been explored to combine the complementary strengths of accurate geometry and rich texture [14], [15]. Hybrid fusion strategies range from colorizing LiDAR point clouds [5] to joint optimization of photometric and geometric constraints [6]. Rigid registration techniques such as ICP [9] remain a standard baseline, with numerous variants addressing outlier rejection, multi-scale matching, and probabilistic correspondences [10]. Nevertheless, cross-modal registration between LiDAR and photogrammetric or NeRF-derived point clouds remains challenging because of differing point densities and incomplete overlap [16], [17], [18]. Previous studies have focused primarily on real-world data, where measurement noise and uncontrolled motion complicate ground-truth evaluation [19], [20], [21].

2.3. Simulation for 3D Perception

Physics-based simulation offers a practical solution for repeatable benchmarking. Platforms such as CARLA [19] and Blender [22] can synthesize photorealistic scenes with configurable sensors (LiDAR, RGB, GPS, IMU) and controllable noise models. These environments allow rigorous testing of perception and mapping algorithms before deployment to the field. In this context, our work leverages CARLA Town 10 to emulate urban scenes and Blender for high-quality image rendering, enabling reproducible LiDAR–photogrammetry fusion experiments and quantitative comparison of reconstruction and registration accuracy.

3. Simulation Testbed and Dataset

Experiments were carried out in CARLA Town 10, a realistic urban environment containing multi-story buildings, road intersections, traffic signs, light poles, benches, bus stops, and playground structures. This setting provides diverse geometric and textural features for evaluating LiDAR–photogrammetry fusion under controlled conditions. A drone-based sensor platform was simulated to acquire aerial data, ensuring wide-area coverage and varied viewpoints. The sensor suite included synchronized RGB, LiDAR, GPS, and IMU modules, mounted on the virtual drone to emulate real-world survey operations.

High-resolution RGB images were rendered using Blender to generate visually consistent inputs for photogrammetry. These images were reconstructed using Nerfstudio (Nerfacto), producing NeRF-based point clouds that represent the photogrammetric perspective of the environment. The resulting NeRF reconstructions were aligned with simulated LiDAR scans through the ICP algorithm, enabling geometric fusion and performance comparison in a shared coordinate frame.

The dataset provides synchronized RGB–LiDAR

sequences, corresponding 6-DoF ground-truth poses, and NeRF reconstructions suitable for both training and validation tasks. These assets collectively allow reproducible analysis of reconstruction fidelity and registration accuracy. Fig. 1 shows representative Town 10 views used for capture, Fig. 2 shows the visualization with the CARLA simulator (point cloud), and Table I summarizes the simulated sensor suite and dataset organization.



Figure 1. Representative views of CARLA Town 10 used for data capture.

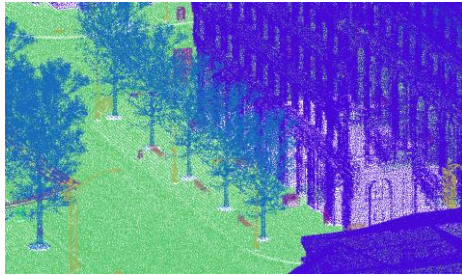


Figure 2. Town 10 point-cloud view used for geometric evaluation.

Table 1. Summary of simulated sensors and dataset composition

Sensor	Modality	Description	Output
RGB Camera	Visual	High-quality images rendered in Blender	Input for NeRF reconstruction
LiDAR	Depth	Simulated CARLA 3D geometry scans	Ground truth for registration
GPS	Position	Provides reference coordinates	Pose alignment
IMU	Motion	Orientation and acceleration	Temporal synchronization

4. Methodology

4.1. Photogrammetry via NeRF

Photogrammetric reconstruction was performed using Nerfstudio (Nerfacto), trained on synthetic RGB images rendered in Blender under controlled lighting. A subset of frames was selected from each drone trajectory to ensure diverse viewpoints and minimal motion blur. Camera poses were derived directly from CARLA's simulated GPS/IMU data and refined through bundle adjustment during NeRF optimization. The complete processing workflow is illustrated in Fig. 3.

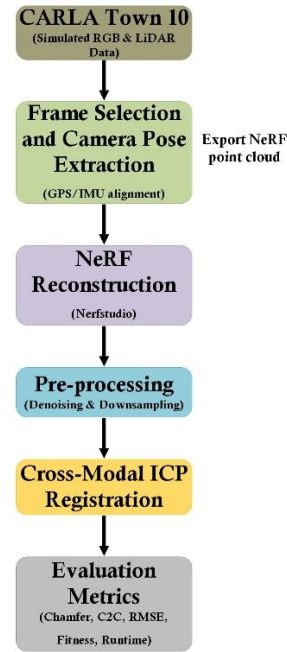


Figure 3. End-to-end LiDAR–Photogrammetry fusion framework.

The NeRF network used the standard volumetric rendering equation as:

$$C(r) = \int_{t_n}^{t_f} T(t) \sigma(r(t)) c(r(t)) dt, \quad (1)$$

where $C(r)$ is the predicted pixel color along a camera ray r , σ denotes volumetric density, and c represents emitted radiance, and $T(t) = \exp\left(-\int_{t_n}^t \sigma(r(s)) ds\right)$ is the accumulated transmittance.

Key configuration parameters included hidden layer size (*hidden_dim* = 256), hash-grid depth (*num_levels* = 32), and adaptive ray sampling. Models were trained for 20,000 iterations using the Adam optimizer ($\text{lr} = 1 \times 10^{-3}$), with exponential decay after 10,000 steps. Denoising and voxel down-sampling were applied to exported point clouds (voxel size = 0.02 m) to suppress background noise before LiDAR alignment. A qualitative example is shown in Fig. 4.

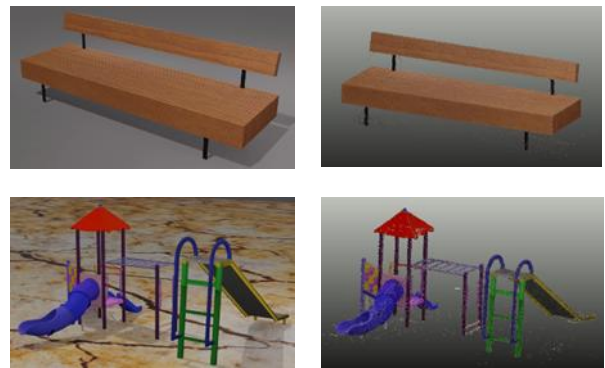


Figure 4. Qualitative NeRF reconstructions (Nerfacto): bench and playground after architecture tuning.

Table 2 lists the NeRF configurations and corresponding training settings.

Table 2. NeRF configurations compared

Config ID	Hidden Dim	Num Levels	Comment
A	128	16	Lower-capacity baseline
B	256	32	Best trade-off (reported)
C	512	32	Higher compute; marginal gain

4.2. Cross-Modal Registration (ICP)

To align NeRF-derived and LiDAR point clouds, we employed the ICP algorithm. Each point cloud was pre-processed by statistical outlier removal and normal estimation. Coarse initialization was obtained using centroids derived from GPS pose priors. ICP minimizes the squared distance between corresponding points:

$$E(R, t) = \frac{1}{N} \sum_{i=1}^N \|p_i - (Rq_i + t)\|^2, \quad (2)$$

where p_i and q_i denote LiDAR and NeRF points, respectively, and (R, t) represent the rotation and translation to be estimated.

A voxel resolution of 0.05 m was used for down-sampling. Maximum correspondence distance was set to 0.1 m, with 10 iterations per run. Convergence was declared when the change in mean-squared error dropped below 10^{-6} m² or pose-update norms fell below 10^{-5} .

4.3. Evaluation Metrics

Reconstruction Quality:

Accuracy between NeRF-reconstructed and LiDAR ground-truth point clouds was measured using the Chamfer distance as:

$$d_{CD}(P, Q) = \frac{1}{|P|} \sum_{p \in P} \min_{q \in Q} \|p - q\|_2^2 + \frac{1}{|Q|} \sum_{q \in Q} \min_{p \in P} \|q - p\|_2^2. \quad (3)$$

and cloud-to-cloud (C2C) distances computed via nearest-neighbor search.

Registration Accuracy:

After ICP, alignment performance was reported using RMSE, fitness score (inlier ratio within correspondence threshold), and runtime per iteration, along with convergence curves.

Runtime and Convergence:

Training and inference times were recorded for each NeRF configuration, while ICP convergence curves were used to analyze stability across object categories (e.g., poles, benches, playgrounds).

5. Experiments

5.1. Ablation on NeRF Capacity and Sampling

We first analyzed how the NeRF network's representational capacity affects reconstruction and

subsequent LiDAR alignment. As described in Section IV-A, the parameters $hidden_dim \in \{32, 64, 128, 256, 512\}$ and $num_levels \in \{32, 64, 128, 256, 512\}$ were varied. Increasing $hidden_dim$ improved structural detail recovery but incurred diminishing returns beyond 256 units. Fig. 5 illustrates this trend, showing that configuration B ($hidden_dim = 256, num_levels = 32$) achieved the lowest Chamfer and C2C distances. This setup also yielded faster and more stable ICP convergence, thus serving as the baseline for subsequent experiments.

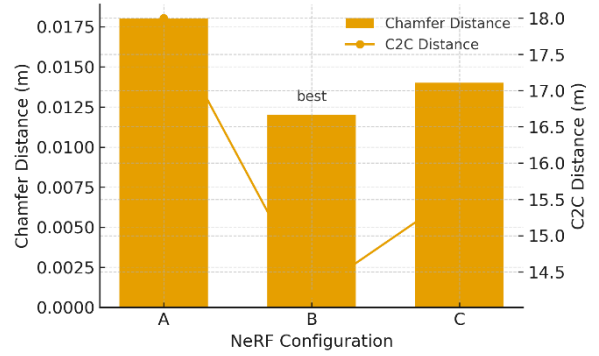


Figure 5. Reconstruction quality vs. NeRF configuration (Chamfer & C2C).

5.2. Object-wise and Scene-wise Registration

For cross-modal registration, five representative urban objects, trash can, pole, playground, road sign, and bench, were evaluated. The NeRF point clouds reconstructed with the optimal configuration (256/32) were aligned with the corresponding LiDAR scans using ICP (10 iterations, voxel = 0.05 m). Table III summarizes the object-wise results, while Fig. 6 presents the ICP convergence curves for “easy” (road sign, pole) and “hard” (playground, bench) objects.

Table 3. Object-wise ICP Results

Object	Initial RMSE (m)	Final RMSE (m)	Best Iter	Fitness (%)	Runtime (s)
Trash can	0.098	0.096	3	97.5	0.92
Pole	0.069	0.069	4	98.2	0.95
Playground	0.200	0.072	7	94.6	1.12
Road sign	0.035	0.034	6	99.1	0.87
Bench	0.150	0.133	4	96.3	0.98

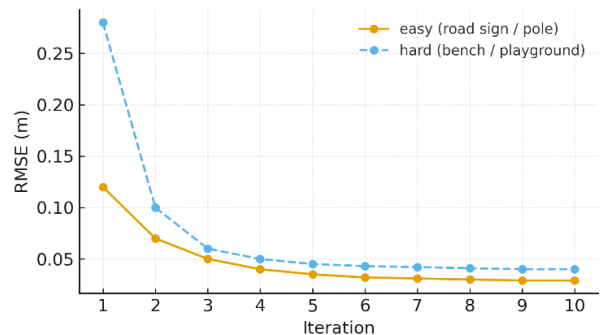


Figure 6. ICP convergence curves (RMSE vs. iteration) for “easy” vs. “hard” objects.

As shown in Fig. 6, structured objects exhibit rapid error

reduction and smooth convergence, whereas complex geometries converge more slowly. Fig. 7(a–e) provides visual overlays of the initial NeRF reconstruction and the final LiDAR–NeRF alignment, highlighting the improved correspondence achieved after ICP registration.

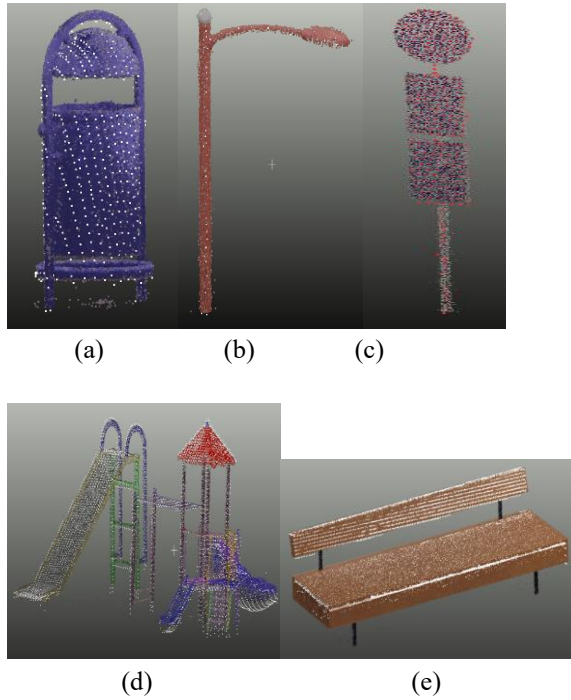


Figure 7. (a–e) Per-object LiDAR–NeRF overlays (trash can, pole, playground, road sign, bench).

5.3. Noise Sensitivity

To evaluate robustness under sensor perturbations, simulated Gaussian noise was applied to LiDAR range and GPS/IMU pose data at three levels (low, medium, high). Table IV summarizes the averaged ICP performance under each condition. At medium noise, RMSE increased by roughly 11%, while high noise caused a 27% degradation and a 2-iteration delay in convergence. Despite this, denoised and voxel-downsampled NeRF inputs maintained > 90% fitness, demonstrating that moderate pre-processing reduces most noise-induced drift.

Table 4. ICP Results under Sensor Noise

Noise Level	RMSE (m)	Fitness (%)	Avg. Iters	Observation
Low	0.071	97.8	5	Stable convergence
Medium	0.079	94.5	6	Minor drift
High	0.090	91.2	7	Slower convergence

6. Discussion and Limitations

The experiments revealed that network capacity and pre-processing quality are the most influential factors for

improving LiDAR–NeRF alignment. Increasing the hidden layer dimension to 256 and adjusting the hash-grid depth yielded significantly lower Chamfer and C2C distances. Normalization and voxel down-sampling before ICP further enhanced convergence, reducing RMSE in complex objects such as playground structures by over 60%. These findings confirm that reconstruction quality directly governs registration stability.

Nevertheless, several failure cases emerged. NeRF reconstructions degraded for textureless or reflective surfaces, such as metal signs, and repetitive geometries like poles often led to local minima during ICP. Fine details, including thin structures, occasionally appeared blurred due to limited sampling density, indicating a need for adaptive ray sampling and improved surface regularization.

A clear speed–accuracy trade-off was observed: higher-capacity NeRFs improved alignment but increased computation time. While suitable for controlled or offline analyses, real-world applications may require lightweight NeRF variants or partial fine-tuning for on-site mapping. Additionally, the current pipeline was evaluated only in a simulated CARLA environment; real-world data would introduce challenges such as illumination variation, occlusion, and sensor drift, requiring domain adaptation or hybrid calibration strategies.

7. Conclusions

This work presented a controlled simulation framework for evaluating LiDAR–photogrammetry fusion using neural radiance fields (NeRF) and iterative closest point (ICP) registration. Through systematic ablations on NeRF capacity and pre-processing, we demonstrated that moderate network scaling and voxel down-sampling significantly improve alignment accuracy and convergence speed. The proposed CARLA-based setup enables reproducible benchmarking of reconstruction and registration quality under varying noise levels and object complexities. Future work will extend this framework to small-scale real-world pilots, emphasizing robust global initialization across large scenes and the integration of uncertainty-aware ICP to handle real sensor variability and dynamic lighting conditions.

Declaration of Ethical Standards

This study is based entirely on simulation-generated data and did not involve human participants, human tissue, animals, or patient-specific records. therefore, ethical approval and informed consent were not required for this work.

CRedit Authorship Contribution Statement

Rytis Maskeliūnas: Conceptualization, Methodology, Supervision, Project administration, Funding, Writing – review & editing. Sarmad Maqsood: Methodology,

Validation, Formal analysis, Investigation, Writing – original draft, Visualization. Ahmad Qurthobi: Data curation, Visualization. Irfan Abbas: Formal analysis. Mantas Vaškevičius: Software, Resources, Data curation, Writing – review & editing. Julius Gelšvartas: Conceptualization, Resources, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING/Acknowledgment

This research project no.: 02-019-K-0044 was funded by the European Union Funds for the period 2021-2027 under the Measure No. 05-001-01-05-07 “Establishing a coherent system for the promotion of innovative activities” under the activity “Stimulating the supply of innovations” under the action “Investing in activities for the development of new high value added products and enabling researchers to participate in R&D activities of enterprises, promotion of intellectual property, early pilot production of new products being developed and preparation for the market” (region of Central and Western Lithuania).

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Author Note

This study was previously presented as a proceedings paper at the *3rd International Conference on Intelligent Systems and New Applications (ICISNA'25)*, held in Antalya, Türkiye, on December 12–14, 2025.

References

- [1] T. Ren and H. Jebelli, “Efficient 3D robotic mapping and navigation method in complex construction environments,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 40, no. 12, pp. 1580-1605, 2025.
- [2] R. Maskeliūnas, S. Maqsood, “Lightweight Attention-Based Framework for Semantic Segmentation and Compression of 3D LiDAR Data”, SETSCI Conference Proceedings, vol. 22, pp. 7-10, 2025.
- [3] P. Chen, X. Zhao, L. Zeng, L. Liu, S. Liu, L. Sun, Z. Li, H. Chen, G. Liu, Z. Qiao, Y. Qu, D. Xu, L. Li, and L. Li, “A Review of Research on SLAM Technology Based on the Fusion of LiDAR and Vision,” *Sensors*, vol. 25, no. 5, 1447, 2025.
- [4] R. Maskeliūnas, S. Maqsood, M. Vaškevičius, and J. Gelšvartas, “Fusing LiDAR and photogrammetry for accurate 3D data: A hybrid approach,” *Remote sensing*, vol. 17, no. 3, pp. 1-27, 2025.
- [5] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, “Robust registration of multimodal remote sensing images based on structural similarity,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2941-2958, 2017.
- [6] K. Istenič, N. Gracias, A. Arnaubec, J. Escartín, and R. Garcia, “Scale accuracy evaluation of image-based 3D reconstruction strategies using laser photogrammetry,” *Remote Sensing*, vol. 11, no. 18, 2093, 2019.
- [7] R. Maskeliūnas, and S. Maqsood, “Hybrid attention-based PTv3-SE model for efficient point cloud segmentation,” *Remote Sensing Applications: Society and Environment*, vol. 41, 101891, 2026.
- [8] M. Gassilloud, B. Koch, and A. Göritz, “Occlusion mapping reveals the impact of flight and sensing parameters on vertical forest structure exploration with cost-effective UAV based laser scanning,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 139, 104493, 2025.
- [9] J. Yang, H. Li, D. Campbell, and Y. Jia, “Go-ICP: A globally optimal solution to 3D ICP point-set registration,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2241-2254, 2015.
- [10] C. H. Lin, W. C. Ma, A. Torralba, and S. Lucey, “Barf: Bundle-adjusting neural radiance fields,” In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5741-5751, 2021.
- [11] Y. Chen and G. H. Lee, “DReg-NeRF: Deep Registration for Neural Radiance Fields,” in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [12] A. T. D. S. Ferreira, C. H. Grohmann, M. C. H. Ribeiro, M. S. T. Santos, R. C. de Oliveira, and E. Siegle, “Beach surface model construction: A strategy approach with structure from motion-multi-view stereo,” *MethodsX*, vol. 12, 102694, 2024.
- [13] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” *Communications of the ACM*, vol. 65, no. 1, pp. 99-106, 2021.
- [14] Y. Li and X. Xiao, “Deep Learning-Based Fusion of Optical, Radar, and LiDAR Data for Advancing Land Monitoring,” *Sensors*, vol. 25, no. 16, 4991, 2025.
- [15] S. Xu, Q. Xue, Z. Chen, S. Fei, and H. Gao, “Complementary information-guided interactive fusion network for HSI and LiDAR data joint classification,” *Expert Systems with Applications*, vol. 298, 129549, 2026.
- [16] N. Xu, R. Qin, and S. Song, “Point cloud registration for LiDAR and photogrammetric data: A critical synthesis and performance analysis on classic and deep learning algorithms,” *ISPRS Open J. Photogramm. Remote Sens.*, vol. 8, p. 100032, 2023.
- [17] K. Ma, F. Yan, S. Li, G. Huang, X. Jia, F. Wang, and L. Chen, “Low-Overlap Registration of Multi-Source LiDAR Point Clouds in Urban Scenes Through Dual-Stage Feature Pruning and Progressive Hierarchical Methods,” *Remote Sensing*, vol. 17, no. 17, 2938, 2025.
- [18] J. Wang and H. Xu, “Cross-modal deep learning framework for 3D reconstruction and information integration of Zhejiang wood carving heritage,” *Scientific Reports*, vol. 16, 465, 2025.
- [19] M. Ramezani, Y. Wang, M. Camurri, D. Wisth, M. Mattamala, and M. Fallon, “The Newer College Dataset: Handheld LiDAR, Inertial and Vision with Ground Truth,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, Oct. 2020, pp. 4353–4360.
- [20] R. Maskeliūnas, S. Maqsood, M. Vaškevičius, and J. Gelšvartas, “Hybrid deep learning and geometric algorithms for individual object detection in urban LiDAR point clouds,” *International Journal of Remote Sensing*, vol. 46, no. 23, pp. 9118–9156, 2025.
- [21] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “CARLA: An open urban driving simulator,” In *Conference on robot learning*, pp. 1-16, 2017.
- [22] T. Roosendaal and Community, *Blender 3.0 User Manual*, Blender Foundation, 2021.

*Research Article***Classification of Orthopedic Diseases Using Biomechanical Properties with Machine Learning Methods****Mehmet Cüneyt ÖZBALCI ^{a,*} , Turgay Tugay BİLGİN ^b** ^a Bursa Technical University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, Bursa, Türkiye.^b Bursa Technical University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, Bursa, Türkiye.

ARTICLE INFO

Article history:

Received 3 February 2026

Accepted 30 March 2026

*Keywords:*Biomechanical Features,
Machine Learning,
Orthopedic Disease
Classification,
SMOTE.

ABSTRACT

Orthopedic diseases significantly affect the musculoskeletal system, reducing patients' functional capacity and quality of life. Accurate and early classification of such conditions is therefore critical for effective clinical decision-making. This study proposes a machine learning-based framework for orthopedic disease classification using biomechanical features, with a particular emphasis on handling class imbalance. A set of classification algorithms, including Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, XGBoost, LightGBM, and a Soft Voting Ensemble (XGBoost + LightGBM), were evaluated on a dataset of 310 patients using 10-fold cross-validation. The impact of the Synthetic Minority Over-sampling Technique (SMOTE) was systematically analyzed by comparing model performance with and without its application. Evaluation metrics included Accuracy, Precision, Recall, F1-score, and macro-average ROC-AUC. Results indicate that addressing class imbalance significantly improves model performance, particularly in terms of ROC-AUC. Among the tested methods, Logistic Regression demonstrated the most stable and competitive results. The best performance was achieved by Logistic Regression with SMOTE, yielding 87% accuracy and a macro-average ROC-AUC of 0.96. These findings highlight the importance of imbalance-aware modeling strategies in orthopedic disease classification.

This is an open access article under the CC BY-SA 4.0 license.
(<https://creativecommons.org/licenses/by-sa/4.0/>)

1. Introduction

Orthopedics is a branch of medical science that studies the musculoskeletal system. It has achieved rapid growth with scientific advances. It examines problems related to bones, muscles, joints, and nerves. Thanks to the rapid development of computer science and technology, significant progress has been made in observing tissue behavior, predicting diseases, and conducting treatment studies. Orthopedic diseases are conditions that affect a large number of people today. As with many diseases, early diagnosis is considered crucial for starting treatment early and achieving a positive response to treatment [1].

Today, a vast amount of data from many different fields is stored and preserved. Existing data is extremely valuable in terms of generating meaningful information,

providing interpretations for various purposes, and offering solutions. Data in the medical world can include patient data, disease data, tissue and organ data, and data related to chemicals in the body. In the field of orthopedics, data related to a person's muscle and bone tissues are also prominent. Based on the data of individuals with diseases, it will be possible to make an early diagnosis for people who may potentially become ill and to start treatment in a timely manner. Thanks to various orthopedic imaging techniques, it is possible to collect clinical data and then make the relevant diagnoses [2]. Biomedical imaging allows for the examination of the inside of the human body and the taking of photographs of bones, muscles, and joints. Diagnostic imaging in orthopedics involves the widespread use of ultrasound, X-

* Corresponding author. E-mail address: mehmet.ozbalci@btu.edu.tr
DOI: 10.58190/ijamec.2026.163

ray, CT (computed tomography) scans, and MRI [3].

Recent studies have increasingly focused on improving the performance of machine learning models for medical and orthopedic disease classification. The problem of class imbalance has also been widely investigated in recent literature. Comparative analyses of machine learning algorithms in medical diagnosis have shown that model performance is highly dependent on both data characteristics and preprocessing strategies.

Elshehewy and Osman [1] proposed an optimization-based approach for orthopedic disease classification and achieved high accuracy by integrating feature selection strategies. Similarly, Zhang et al. [2] developed a hybrid SMOTE-RFE-XGBoost model for spinal disease classification, demonstrating that combining feature selection with imbalance handling techniques can significantly improve predictive performance. Kivrak et al. [4] evaluated multiple machine learning algorithms using SMOTE and reported notable improvements in recall and F1-score for minority classes. Yang et al. [5] proposed a hybrid framework combining SMOTE with ensemble learning methods such as LightGBM and Random Forest, achieving improved classification performance in healthcare applications. Gyasi-Agyei [6] demonstrated that no single model consistently outperforms others across all scenarios, highlighting the importance of dataset-specific evaluation. In orthopedic-related applications, Rezapour et al. [7] applied machine learning techniques with SMOTE to gait analysis data and showed that imbalance-aware learning improves classification robustness.

In the study by Hafsa Binte Kibria and Abdul Matin, fusion models were created to diagnose and evaluate the severity of CVDs (cardiovascular diseases). Artificial neural networks, SVM, logistic regression, decision trees, Random Forest, and AdaBoost algorithms were applied to the heart disease dataset to perform disease prediction. To balance the imbalanced dataset, Randomoversampler was used only for multi-class classification, and a weighted score fusion approach was used to improve classification performance. In the proposed approach, the highest accuracy for multi-class classification was found to be 75.41% with logistic regression and artificial neural networks, while for binary classification, it was 95.08% with AdaBoost and decision trees [8]. In the study conducted by Shenglong Li and Xiaojing Zhang, an orthopedic assistant classification prediction model developed with the XGBoost algorithm is proposed. The model was created using the Random Forest algorithm, XGBoost, and the associated classification algorithm, respectively. The results obtained show that the XGBoost algorithm has a higher accuracy with a value of approximately 95% [9]. The study conducted by Cai-Jin Ling and his team investigated the effectiveness of biomedical sensors and medical image segmentation

algorithms in the diagnosis and treatment of orthopedic diseases. This study aims to develop image segmentation methods by collecting data from MR and scan images and then using machine learning techniques. The data was divided into two groups. The first group is a control group that attempts to diagnose orthopedic diseases using X-ray images with traditional methods, while the second group is an experimental group that uses machine learning techniques with MR or CT scanning methods to diagnose orthopedic diseases. The proposed method was evaluated using standard performance metrics such as VAS score, Ramsay score, disease classification, treatment effect, and observation index. In the study, CNN, KNN, Fast KNN, and Mask R-CNN machine learning methods were applied and their performance was compared. Looking at all the results obtained, Mask R-CNN stands out as the method that achieved the highest performance [10].

Nadia Rubaiyat and colleagues aims to predict orthopedic diseases early. To this end, three different machine learning algorithms—logistic regression, random forest, and KNN—were applied to a dataset of 310 patients containing six biomechanical features describing the pelvic and lumbar conditions of the patients. When the results were compared, the random forest algorithm yielded the best result with an accuracy rate of 89%. KNN achieved 85% accuracy, while logistic regression achieved 81% accuracy [11]. In the study conducted by Nasrin Jahan and colleagues, a dataset containing 310 examples was used to classify patients using various machine learning algorithms. Classification was performed in two stages. In the two-stage classification, patients were classified as normal or abnormal in the first stage, and then classified according to their physical condition in the second stage. Performance metrics were calculated along with the performance rates. The KNN algorithm yielded the highest accuracy rate. Most other classification algorithms also achieved successful results with an accuracy above 80% [12]. In the study conducted by Kamrul Hasan and colleagues, an attempt was made to predict the presence or absence of orthopedic disease. The classification performance of various classification algorithms was measured. Each patient in the dataset is represented by six biomechanical features derived from the shape and orientation of the pelvis and lumbar spine. While most of the applied algorithms achieved over 90% accuracy, the decision tree (DT) algorithm achieved 99% accuracy, showing a significant difference [13]. In the study conducted by Qiuchong Chen and colleagues, AKI risk prediction in elderly patients after orthopedic surgery is performed based on machine learning algorithm models. Nine different machine learning algorithms were applied to an experimental group of 1000 postoperative AKI patients who underwent orthopedic surgery. Intraoperative information and preoperative clinical characteristics were used in the prediction phase. The ideal model was

evaluated by calculating AUC, sensitivity, specificity, and accuracy. In this way, the most suitable model was selected. Logistic regression yielded the best results [14].

In the study conducted by Mantzaris and colleagues, various algorithms were evaluated in terms of classification performance for osteoporosis, an orthopedic disease. The aim of the study was to assist experts in predicting osteoporosis by avoiding unnecessary further tests using bone densitometry. Multi-layer perceptrons (MLP) and probabilistic neural networks (PNN), which are feedforward networks, were used to predict osteoporosis risk factors. PNNs were applied with propagation values ranging from 0.1 to 50 and with 4 or 2 neurons in the output layer, depending on the coding of the desired outcome of osteoporosis. Looking at the experimental results, it was observed that PNNs performed better. In addition, as the spread values increased, the overfitting problem was more frequently observed in MLPs [15]. In the study conducted by Seok Won Chung and colleagues, a study was performed to detect and classify proximal humerus fractures using CNN with plain anterior-posterior shoulder radiographs. The results were evaluated by measuring AUC, sensitivity, specificity, and Youden index. To distinguish normal shoulders from proximal humerus fractures, 96% accuracy, 1.00 AUC, 0.99 sensitivity, 0.97 specificity, and 0.97 Youden index were obtained. To classify the fracture type, 65-86% accuracy, 0.90-0.98 AUC, 0.88-0.97 sensitivity, 0.83-0.94 specificity, and 0.71-0.90 Youden index were obtained [16]. Jakub OLCZAK and colleagues classified ankle fractures. A neural network based on the ResNet architecture was trained using 4,941 radiographic ankle examination results. All images were classified according to the AO/OTA (AO Foundation/Orthopedic Trauma Association) 2018 classification, and the network performance was evaluated against a test set of 400 patients independently reviewed by two expert observers. The results showed that the AUC value reached up to 0.93 for type B fractures. Type A fractures showed the worst performance [17]. This study investigated the efficient detection of regions of interest in hip MRI and the accurate identification of the femoral head and necrotic region. The study, conducted with 484 patients, utilized 3,937 labeled hip MRI images. Lesion detection achieved a mAP value of up to 94.7%. Cao and colleagues aimed to efficiently identify the region of interest in hip MRI and accurately define the femoral head and necrotic region. Comparative analyses were performed using Swin-Unet, U-net, and YOLO models. The study, conducted with 484 patients, utilized 3,937 labeled hip MRI images. The lesion detection mAP value reached up to 94.7% [18]. A study on the classification of orthopedic diseases was conducted by Elshevey and Osman. For feature selection in the study, binary particle swarm optimization (BPSO), binary gray wolf optimization (BGWO), and binary whale

optimization algorithm (BWAO) were used. An average error rate reduction of 47.29% was observed in the breadth-first search method. The highest accuracy rate was achieved with BFS-Random Forest, with 99.41% accuracy [19].

Although there are numerous studies in the literature using the same dataset, studies that systematically compare the effect of SMOTE on different classifiers in terms of macro-average metrics are limited. This study aims to fill this gap.

1.1. Key Contributions of This Study

A comprehensive machine learning framework is proposed for the classification of orthopedic diseases using biomechanical features, addressing a clinically relevant multi-class problem. The effect of class imbalance is systematically investigated by integrating the Synthetic Minority Over-sampling Technique (SMOTE) into the modeling pipeline and comparing results with and without its application. A wide range of machine learning algorithms, including Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, XGBoost, LightGBM, and a Soft Voting Ensemble model, are evaluated under a unified experimental setup using stratified 10-fold cross-validation.

Model performance is assessed using macro-average evaluation metrics (Accuracy, Precision, Recall, F1-score, and ROC-AUC), providing a balanced and reliable comparison across imbalanced class distributions. Experimental results demonstrate that Logistic Regression, particularly when combined with SMOTE, achieves the most stable and competitive performance, offering practical insights for clinical decision support systems.

The study provides a systematic and comparative analysis of imbalance-aware learning strategies, contributing to the limited literature on SMOTE-based evaluation using macro-average metrics in orthopedic disease classification.

2. Material and Methods

The steps of the study are shown step by step in Figure 1 below.

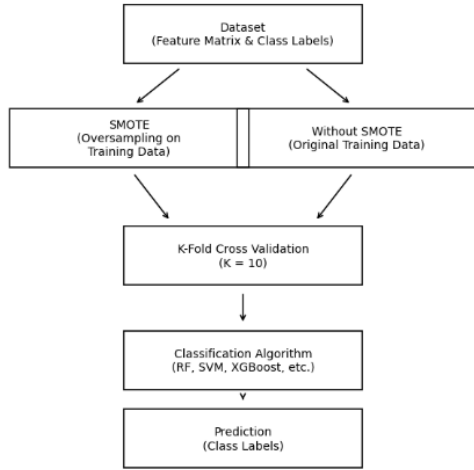


Figure 1. Phases of Study

Results were obtained and compared when SMOTE was applied and when it was not applied during this process. SMOTE is applied only to the training data, followed by classification using 10-fold cross-validation and multiple machine learning algorithms. The hyperparameters used in all models were selected from the default values commonly used in the literature. Complex hyperparameter optimization was not applied in order to prevent overfitting. Finally, performance metrics are obtained.

2.1. Dataset

The dataset [20] used contains the biomechanical characteristics that define the pelvic and lumbar conditions of orthopedic patients and, if a disease condition is present, the name of that disease. These biomechanical characteristics consist of six features: pelvic incidence, pelvic tilt numeric, lumbar lordosis angle, sacral slope, pelvic radius, and grade of spondyloisthesis. The condition of each of these characteristics triggering the disease has been examined. It investigates whether 310 different individuals have one of two different diseases or neither of these two diseases, i.e., whether they are healthy. All six different characteristics of each individual are present in the dataset. The diseases are spondylolisthesis and hernia. In addition to these two diseases, there is also a class designated as healthy, or normal. Thus, there are three different classes in total. Of these three classes, 60 belong to the hernia class, 150 to the spondylolisthesis class, and 100 to the normal class. The data is the Biomechanical features of orthopedic patients dataset obtained from the Kaggle website. The dataset used in this study is publicly available and does not contain any personally identifiable information; therefore, no ethics committee approval is required. The dataset and its characteristics are provided in Table 1 below.

Table 1. Description of Biomechanical Features Used in the Dataset

Number	Feature	Definition
1	Pelvic incidence	It refers to the angle between the line perpendicular to the sacral plate at its midpoint and the line connecting this point to the femoral head axis.
2	Pelvic tilt	It refers to the angle formed by the vertical line passing through the femoral head and the line connecting the center of the sacral plane.
3	Lumbar lordosis angle	It refers to the angle formed by the inward curvature of the lumbar spine region located just above the hips.
4	Sacral slope	It refers to the angle between the line tangent to the sacral plateau and the vertical line passing through the exact center point of the sacral plateau.
5	Pelvic radius	It refers to the distance from the hip axis to the posterior-superior corner of the S1 endplate.
6	Grade of spondylolisthesis	It indicates the degree of spinal displacement.

To ensure that model performance can be evaluated in a reliable and generalizable manner, the Stratified K-Fold Cross Validation method was used in this study. The dataset was divided into 10 sub-sets (folds) while preserving the class distribution; in each iteration, one sub-set was used as test data while the remaining nine sub-sets were evaluated during the training phase. This minimized bias in random data division. Due to the limited size of the dataset, 10 folds were deemed appropriate.

2.2. Synthetic Minority over Sampling Technique (SMOTE)

The primary goal of SMOTE is to increase the number of examples belonging to minority classes while preserving the interpolation of these examples in the feature space. This approach aims to improve the model's generalization ability by reducing class imbalance [21]. In the SMOTE method, synthetic samples are generated using linear interpolation points between minority class samples. This makes the dataset relatively balanced. As a result, it provides balanced and reliable classification performance for minority classes [22].

The Synthetic Minority Over-sampling Technique (SMOTE) was used to reduce the potential negative effects of class imbalance on classification performance. SMOTE generates synthetic data by creating linear combinations of existing samples for minority classes [23]. This relatively balances the number of samples between classes. This method particularly improves macro-average performance metrics. Furthermore, it has a positive effect on the model's generalization ability on minority classes.

The SMOTE process was applied only on the training subset in each cross-validation iteration. Test data were not

included in any synthetic sampling process.

2.3. Random Forest

The Random Forest algorithm is a flexible machine learning algorithm that can be applied for various purposes. Random Forest can provide solutions for classification and regression problems. It appears as an advanced derivative of decision trees [24]. In this algorithm, partitioning is applied randomly. The goal is to find the optimal branch. Sub-branches are then created based on this [25].

Random Forest is an ensemble learning algorithm. The main goal in ensemble learning methods is to improve performance by combining models. In Random Forest, continuous and non-continuous variables can be used together in large datasets [26].

2.4. Support Vector Machines

SVM is an effective method that produces successful results and is commonly used for classification. It enables the prediction of labels using one or more feature vectors [27]. The classification process is performed by creating a decision boundary between classes [28]. There are three commonly used kernel functions for classifiers: the radial basis kernel, linear kernel, and polynomial kernel function. In addition, the sigmoid kernel function can also be used in some cases. While data preprocessing and similar operations performed beforehand are of great importance, the kernel function applied also has a significant impact on performance [29]. The RBF kernel function, which yielded the most successful results in this study, was used. The equation of RBF given below.

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (1)$$

In the above equation, x_i, x_j represents the data points, while $\|x_i - x_j\|^2$ is the square of the Euclidean distance. Here γ is the spread parameter of the kernel.

Data classification is performed using hyperplanes. The best classifier is the most ideal hyperplane. The distance between hyperplanes is called the margin [30]. The small number of operations during the learning phase in the SVM algorithm positively affects performance.

2.5. Naive Bayes Classifier

Naive Bayes is a powerful classification algorithm widely used in classification problems and based on Bayes' theorem. In this method, which is shaped on the basis of conditional probabilities, the effect and dominance of all attributes are investigated [31]. This method is based on selecting the option with the highest probability [32]. The classification process is performed by combining the effects of different features on the result [33]. To achieve the optimal result, the final calculation is made based on the combined probabilities of different situations. The formula of Bayes' theorem used in calculating the combined probability is as follows.

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \quad (2)$$

In the above equation C_k represents class, while x represents feature vector.

2.6. Logistic Regression

Logistic regression is a statistical method widely used in predictive analysis. It was developed to model the relationship between one or more independent variables and a dependent variable, and it is particularly effective in linear classification problems [34].

Logistic regression is preferred because it offers a more interpretable structure compared to other multivariate analysis methods. Thanks to this flexibility, it is an effective method that can provide reliable and successful classification performance.

2.7. XGBoost

Extreme Gradient Boosting (XGBoost) is a high-performance machine learning algorithm based on the gradient boosting approach. It works by sequentially training decision trees to minimize the errors of previous models. XGBoost also demonstrates successful performance in parallel processing [35].

XGBoost relatively prevents overfitting thanks to its regularization mechanism. Performance is also improved through parallel computing and efficient memory usage. This makes it a classification method that stands out for its successful performance, especially on imbalanced datasets.

2.8. LightGBM

Light Gradient Boosting Machine (LightGBM) has a structure that uses gradient boosting-based decision trees. It is a machine learning algorithm that stands out for its high speed and efficient memory consumption [36]. LightGBM is successful in learning complex situations with its leaf-wise tree growth structure. The histogram-based learning approach reduces computational costs. LightGBM has become a popular method, especially for high-dimensional datasets.

2.9. XGBoost + LightGBM Soft Voting Ensemble

XGBoost + LightGBM Soft Voting Ensemble is an ensemble method based on combining the outputs of gradient boosting-based XGBoost and LightGBM algorithms. In the soft voting strategy, classification is performed using probabilities obtained by weighted averaging and combining the class probabilities produced by each base model. Combining the strengths of both algorithms, this method leverages XGBoost's regularization capabilities and LightGBM's fast learning ability to improve the model's classification performance.

2.10. Statistical Analysis

In this study, accuracy, recall, precision, specificity, and

F1-Score [37] were computed for a three-class classification problem using confusion matrices.

TP: Correct prediction of the positive class.

TN: Correct prediction of the negative class.

FP: Incorrect prediction of the positive class.

FN: Incorrect prediction of the negative class.

2.10.1. Accuracy

Accuracy represents the overall accuracy of the model by measuring the proportion of correctly classified examples out of all examples.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

2.10.2. Recall

Recall, known as sensitivity or true positive rate, measures a model's ability to correctly identify positive examples. It is the ratio of true positives to the total of true positives and false negatives.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

2.10.3. Precision

Precision measures the proportion of correctly predicted positive examples in all positively predicted examples.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

2.10.4. Specificity

Specificity, known as the true negative rate, measures the proportion of correctly identified negative examples.

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (6)$$

2.10.5. F1 Score

The F1 score is calculated as the harmonic mean of precision and recall rates, providing a balanced measure of both metrics.

$$\text{F1 Score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (7)$$

3. Results and Discussion

The performance of the Random Forest classifier was evaluated using macro-average Receiver Operating Characteristic (ROC) analysis, confusion matrices, and cumulative Area Under the Curve (AUC) scores.

Macro-average ROC-AUC was preferred in this study due to the imbalanced class distribution in the dataset. Micro-averaging aggregates the contributions of all samples and is therefore dominated by majority classes, potentially masking the performance on minority classes. In contrast, macro-averaging computes the ROC-AUC independently for each class and then averages the results, ensuring equal importance is given to all classes. This makes it a more appropriate and reliable metric for evaluating model performance in imbalanced multi-class

classification problems.

3.1. Results obtained with Random Forest

An average AUC value of 0.949 ± 0.029 was obtained from the macro-average ROC curve. The average confusion matrix showed balanced classification performance between classes. Feature importance analysis based on impurity reduction revealed that the degree of spondylolisthesis was the most influential feature on classification.

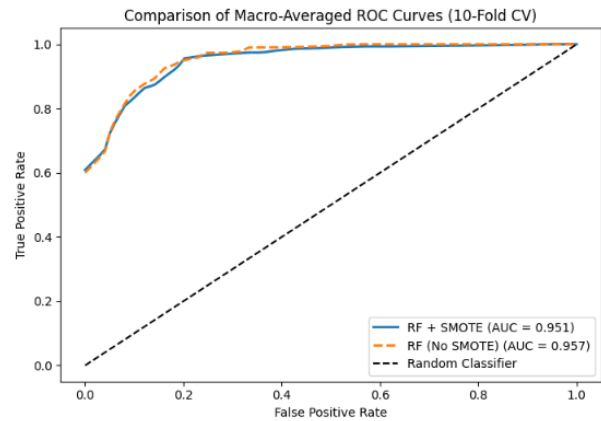


Figure 2. Macro-Averaged ROC Curves of Random Forest: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in the Figure 3, the Random Forest classifier has high discriminative power in both cases. The fact that the AUC value obtained without applying SMOTE (0.957) is only slightly higher than the value obtained with SMOTE (0.951) indicates that class imbalance in this dataset did not significantly negatively affect RF performance.

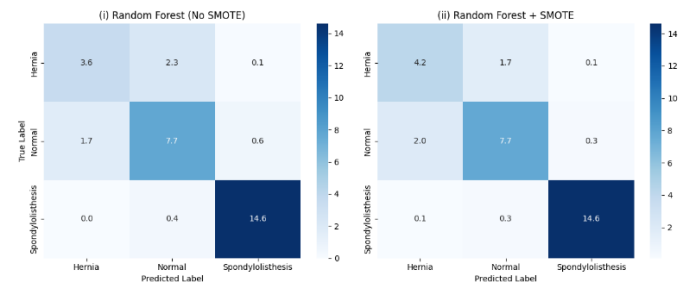


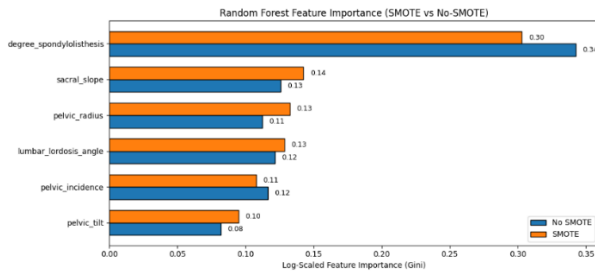
Figure 3. Average confusion matrices of Random Forest: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 4, the application of SMOTE has had a slightly positive effect on the Hernia class. No significant difference was observed for the other two classes. Looking at the overall situation in the confusion matrix, a very slight improvement was observed with SMOTE.

Table 2. Metrics obtained with Random Forest (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.85	0.06	0.84	0.06
2	Precision (Macro)	0.83	0.08	0.81	0.08
3	Recall (Macro)	0.81	0.07	0.78	0.06
4	F1-Score (Macro)	0.81	0.07	0.77	0.07
5	ROC-AUC (Macro, OvR)	0.95	0.04	0.95	0.03

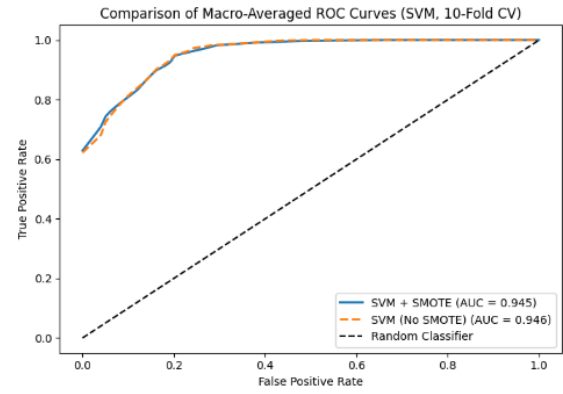
As shown in Table 2, the SMOTE application provides consistent improvements in macro-average Precision, Recall, and F1 score, showing an increase of approximately 2-4% compared to the setting without SMOTE. The macro-average Recall increased from 0.78 to 0.81, indicating increased sensitivity towards minority classes. While the overall Accuracy remained similar in both settings (0.85 and 0.84), the SMOTE-based model showed higher Precision, Recall, and F1 scores. The macro-average ROC-AUC was high and close in both cases (0.95). This indicates that the nature of class separability is inherently strong.

**Figure 4.** Comparison of Random Forest feature importance scores with and without SMOTE

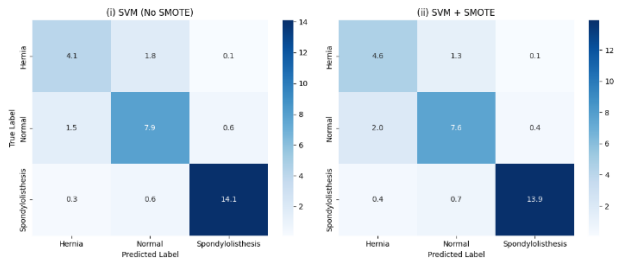
In Figure 5, the feature importance analysis performed using the Random Forest model shows the importance of features when SMOTE is used and when it is not used. In both cases, the degree of spondylolisthesis emerges as the most dominant feature. This is followed by sacral tilt and lumbar lordosis angle. Although there are minor differences in the relative importance values between the two configurations, the overall feature hierarchy is similar. This indicates that the application of SMOTE does not alter the fundamental decision structure of the model.

3.2. Results obtained with SVM

The macro average ROC curve yielded an average AUC value of approximately 0.95 both when SMOTE was used and when SMOTE was not used. This value indicates high discrimination power between classes. The average confusion matrices showed stable and balanced classification performance with limited confusion between classes. The results in SVM demonstrate that the SVM model provides robust and consistent performance both with and without SMOTE.

**Figure 5.** Macro-Averaged ROC Curves of SVM: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 6, the SVM model exhibits very similar macro-average ROC-AUC values when SMOTE is applied and when it is not applied ($AUC \approx 0.945-0.946$). This indicates that SVM is relatively robust to class imbalance in the dataset. Both ROC curves remain well above the random baseline at all false positive rates.

**Figure 6.** Average confusion matrices of SVM: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 7, the confusion matrices obtained from the SVM classifier indicate that the model achieves high true positive rates for the Spondylolisthesis class under both conditions. With the inclusion of SMOTE, a slight improvement is achieved in the correct classification of the Hernia class, while there is a very slight deterioration for the Normal and Spondylolisthesis classes. These results show that the results in SVM are close to each other when SMOTE is included and when it is not included.

Table 3. Metrics obtained with SVM (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.84	0.09	0.84	0.08
2	Precision (Macro)	0.82	0.10	0.82	0.10
3	Recall (Macro)	0.82	0.09	0.80	0.09
4	F1-Score (Macro)	0.81	0.10	0.81	0.10
5	ROC-AUC (Macro, OvR)	0.95	0.03	0.95	0.03

As shown in Table 3, the SVM model performs well both with and without SMOTE. While accuracy, precision, F1 score, and ROC-AUC values remain nearly identical in both settings, a slight improvement in the macro-average recall value is observed when SMOTE is applied.

3.3. Results obtained with Naive Bayes

The macro average ROC analysis yielded an average AUC value of approximately 0.95. These values indicate that the model performs well in distinguishing between classes. The average confusion matrices showed consistent classification performance across all classes. They yielded successful classification rates. Additionally, the results obtained without using SMOTE showed only a limited improvement in the performance of the Naive Bayes classifier.

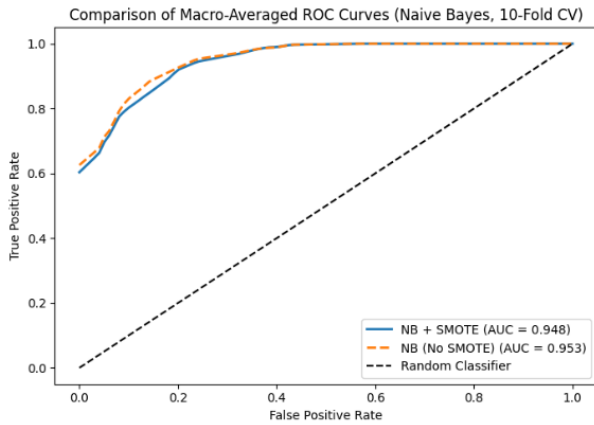


Figure 7. Macro-Averaged ROC Curves of Naive Bayes: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 8, the Naive Bayes model achieves a slightly higher value than the macro-average ROC-AUC value obtained when SMOTE is not used. When SMOTE was not used, the macro average AUC value was slightly higher. Overall, both ROC curves yielded successful results and remained well above the random baseline for all false positive rates.

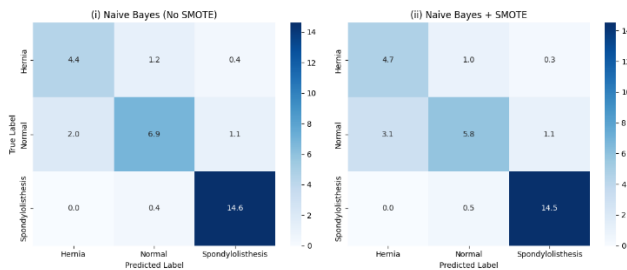


Figure 8. Average confusion matrices of Naive Bayes: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 9, the confusion matrices obtained from the Naive Bayes classifier indicate that the model achieves high true positive rates for the Spondylolisthesis class under both conditions. When SMOTE is applied, the classification performance for the Hernia and Normal classes is higher compared to when SMOTE is not applied. In particular, a significant difference of 1.1 points is observed for the Normal class. Looking at the overall results, it is seen that when SMOTE is not used, the Naive Bayes classifier successfully distinguishes all classes.

Table 4. Metrics obtained with Naive Bayes (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.81	0.05	0.84	0.07
2	Precision (Macro)	0.79	0.07	0.82	0.09
3	Recall (Macro)	0.78	0.06	0.80	0.08
4	F1-Score (Macro)	0.77	0.07	0.80	0.09
5	ROC-AUC (Macro, OvR)	0.95	0.03	0.95	0.03

As shown in Table 4, the Naive Bayes classifier performs better when SMOTE is not applied. Higher values were obtained without SMOTE for all metrics.

3.4. Results obtained with Logistic Regression

The macro average ROC analysis yielded an average AUC value of approximately 0.96. These values indicate that the model performed well in distinguishing between classes. The average confusion matrices showed good classification performance across all classes. Looking at the overall situation, the results obtained when SMOTE was used revealed that the Logistic Regression classifier performed more successfully.

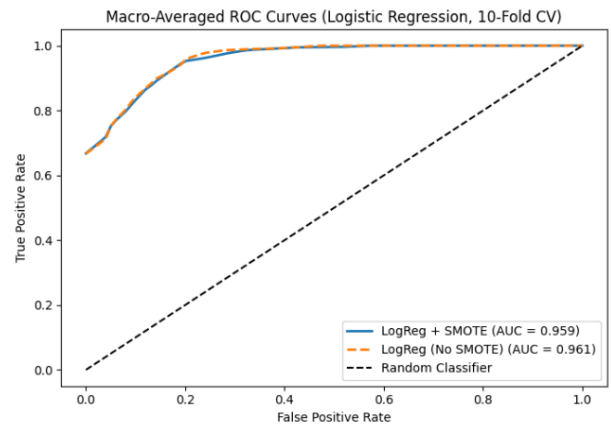


Figure 9. Macro-Averaged ROC Curves of Logistic Regression: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in the Figure 10, the macro-average AUC values are quite close to each other when SMOTE is applied and when it is not applied. Logistic regression performed quite well with an AUC value of approximately 0.96. These results show that logistic regression produced quite successful results in both SMOTE and non-SMOTE conditions.

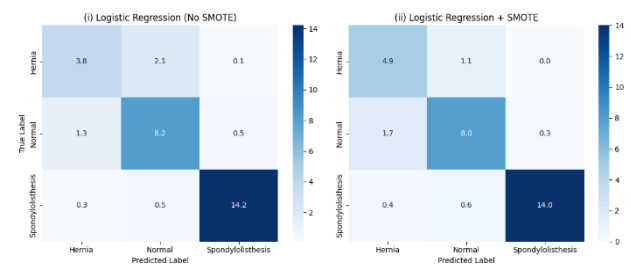


Figure 10. Average confusion matrices of Logistic Regression: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 11, when applied with the SMOTE logistic regression model, a small increase is observed in the accuracy of the Hernia class, while a very slight decrease is observed in the Spondylolisthesis and Normal classes.

Table 5. Metrics obtained with Logistic Regression (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.87	0.07	0.85	0.07
2	Precision (Macro)	0.85	0.08	0.83	0.10
3	Recall (Macro)	0.85	0.08	0.80	0.08
4	F1-Score (Macro)	0.84	0.08	0.80	0.09
5	ROC-AUC (Macro, OvR)	0.96	0.03	0.96	0.03

Table 5 shows the performance metrics obtained with and without SMOTE applied to the logistic regression model, indicating that SMOTE yielded more successful results in all metrics except ROC-AUC. With the use of SMOTE, a slight increase was observed in accuracy, macro-average precision, recall, and F1-score values. The ROC-AUC value was 0.96 in both cases. This indicates that the model has a high ability to distinguish between classes. It appears that SMOTE improves the balance between classes rather than its separation power.

3.5. Results obtained with XGBoost

An average AUC value of approximately 0.95 was obtained from the ROC curve when XGBoost was used with and without SMOTE. The average confusion matrix showed balanced classification performance between classes. Positive effects of applying SMOTE were observed in the performance metrics.

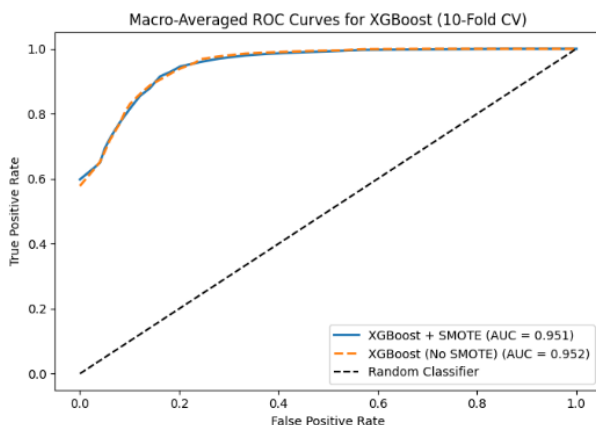


Figure 11. Macro-Averaged ROC Curves of XGBoost: (i) no SMOTE, (ii) SMOTE (10-fold CV)

The macro-average ROC curves presented in Figure 12 show that the XGBoost model achieves high AUC values both with and without SMOTE. In both cases, the AUC is approximately 0.95, and the fact that the curves are significantly above the random classifier line indicates that the model has a high level of ability to distinguish between classes.

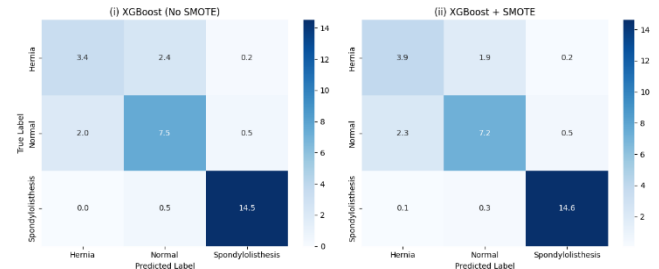


Figure 12. Average confusion matrices of XGBoost: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As seen in the Figure 13, the classification performance of the XGBoost model before and after applying SMOTE is similar. Before SMOTE, the model was more successful in distinguishing the Hernia class, but relatively less successful in the Spondylolisthesis and Normal classes.

Table 6. Metrics obtained with XGBoost (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.83	0.05	0.82	0.06
2	Precision (Macro)	0.80	0.07	0.78	0.07
3	Recall (Macro)	0.78	0.06	0.76	0.06
4	F1-Score (Macro)	0.78	0.06	0.76	0.07
5	ROC-AUC (Macro, OvR)	0.95	0.03	0.95	0.03

Table 6 shows that the XGBoost model using SMOTE performed slightly better than the ROC-AUC for all metrics. The AUC value remaining stable at approximately 0.95 in both cases indicates that the model successfully distinguishes between classes.

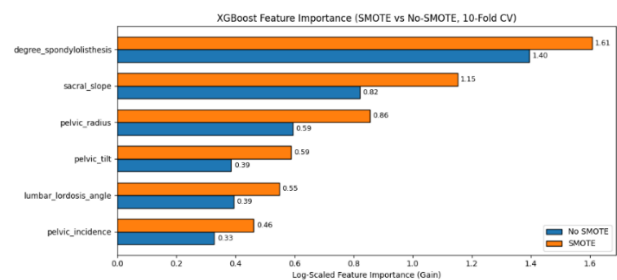


Figure 13. Comparison of XGBoost feature importance scores with and without SMOTE

Figure 14 shows that the feature importance analysis performed using the XGBoost model indicates the importance of features when SMOTE is used and when it is not used. In both cases, the degree of spondylolisthesis was observed to be the most dominant feature. This was followed by sacral tilt and lumbar lordosis angle. It was observed that the application of SMOTE did not change the basic decision structure of the model.

3.6. Results obtained with LightGBM

When LightGBM was used with and without SMOTE, an average AUC value of approximately 0.94 was obtained from the ROC curve. The average confusion matrix

showed similar performance in both cases. A slight positive effect of applying SMOTE was observed in the performance metrics.

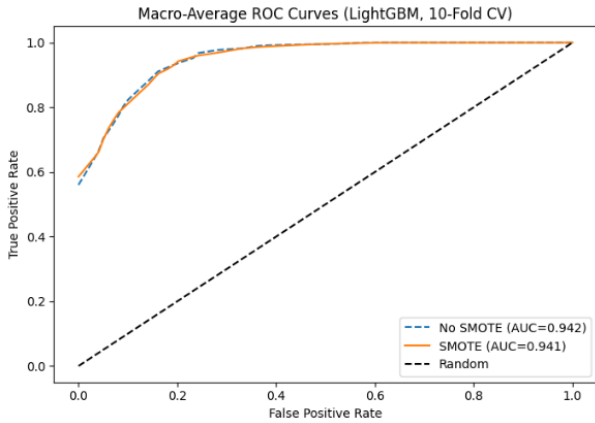


Figure 14. Macro-Averaged ROC Curves of LightGBM: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 15, the LightGBM model achieves an AUC value of approximately 0.94 with and without SMOTE applied, indicating nearly identical performance. Thus, the model demonstrates successful classification performance in both scenarios.

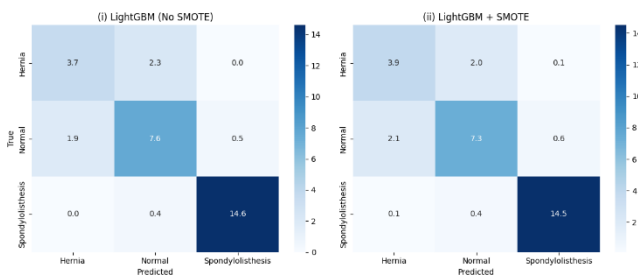


Figure 15. Average confusion matrices of LightGBM: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As seen in the Figure 16, the confusion matrices of the LightGBM model before and after applying SMOTE show similar values for all classes. After applying SMOTE, a slight improvement is observed in the Hernia class. Before applying SMOTE, the performance of the Normal and Spondylolisthesis classes was slightly higher.

Table 7. Metrics obtained with LightGBM (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.84	0.04	0.83	0.05
2	Precision (Macro)	0.81	0.06	0.80	0.08
3	Recall (Macro)	0.78	0.05	0.78	0.07
4	F1-Score (Macro)	0.78	0.06	0.78	0.07
5	ROC-AUC (Macro, OvR)	0.94	0.02	0.94	0.03

As shown in Table 7, applying SMOTE results in a slight increase in the Accuracy and Precision (Macro) metrics, while the Recall, F1-Score, and ROC-AUC values remain nearly the same.

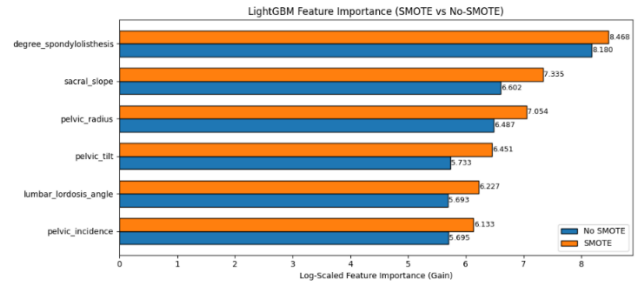


Figure 16. Comparison of LightGBM feature importance scores with and without SMOTE

Figure 17 shows the feature importance analysis performed using the LightGBM model, indicating the importance of features when SMOTE is used and when it is not used. In both cases, the degree of spondylolisthesis was observed to be the most dominant feature. This was followed by sacral tilt and lumbar lordosis angle. The other features were observed to have similar values.

3.7. Results obtained with XGBoost + LightGBM Soft Voting Ensemble

When XGBoost + LightGBM Soft Voting Ensemble was used with and without SMOTE, an average AUC value of approximately 0.94 was obtained from the ROC curve. The average confusion matrix showed similar performance in both cases. No significant positive effect has been observed in the performance metrics of the SMOTE application.

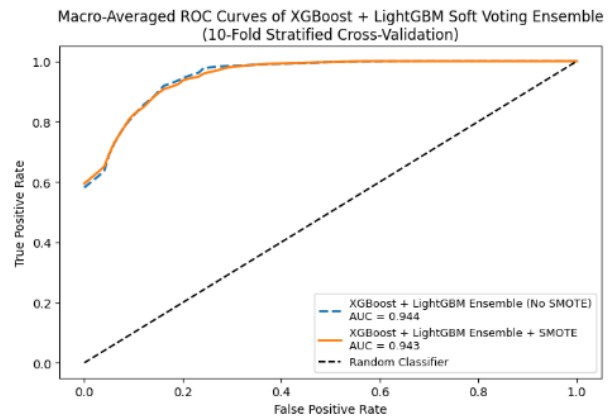


Figure 17. Macro-Averaged ROC Curves of XGBoost + LightGBM Soft Voting Ensemble: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 18, when SMOTE is applied with the XGBoost + LightGBM soft voting ensemble model, the AUC values obtained are approximately 0.94, which is almost the same as when SMOTE is not applied. It can be seen that the AUC value produced by the ensemble structure is close to the average of both algorithms.

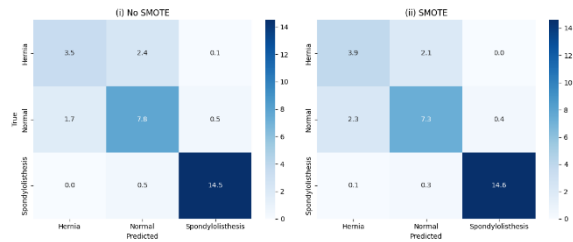


Figure 18. Average confusion matrices of XGBoost + LightGBM Soft Voting Ensemble: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in the Figure 19, the confusion matrices of the XGBoost + LightGBM Soft Voting Ensemble model before and after applying SMOTE show close values for all classes. After applying SMOTE, a slight improvement is observed in the Hernia class. Before applying SMOTE, the performance of the Normal and Spondylolisthesis classes was slightly higher.

Table 8. Metrics obtained with XGBoost + LightGBM Soft Voting Ensemble (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.83	0.06	0.83	0.05
2	Precision (Macro)	0.80	0.08	0.79	0.07
3	Recall (Macro)	0.78	0.07	0.78	0.06
4	F1-Score (Macro)	0.78	0.08	0.78	0.06
5	ROC-AUC (Macro, OvR)	0.94	0.03	0.94	0.03

As shown in Table 8, the XGBoost + LightGBM Soft Voting Ensemble model produced very similar values both when SMOTE was applied and when it was not applied. Only a slight improvement in Precision (Macro) was achieved with SMOTE.

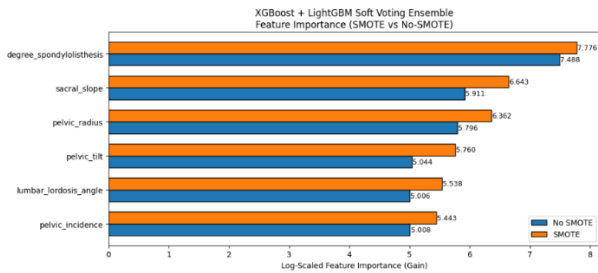


Figure 19. Comparison of XGBoost + LightGBM feature importance scores with and without SMOTE

Figure 20 shows the feature importance analysis performed using the XGBoost + LightGBM model, indicating the importance of features when SMOTE is used and when it is not used. In both cases, the degree of spondylolisthesis was again observed as the most dominant feature. This was followed by sacral tilt and lumbar lordosis angle. Other features also showed similar values.

3.8. Overall Results

The results obtained for all classifiers using both SMOTE and without SMOTE in the experiments conducted are presented in Table 9 below.

Table 9. Performance comparison of classifiers with and without SMOTE (Stratified 10-Fold CV)

Classifier	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	ROC-AUC
Random Forest+SMOTE	0.85±0.06	0.83±0.08	0.81±0.07	0.81±0.07	0.95±0.04
Random Forest	0.84±0.06	0.81±0.08	0.78±0.06	0.77±0.07	0.95±0.03
SVM+SMOTE	0.84±0.09	0.82±0.10	0.82±0.09	0.81±0.10	0.95±0.03
SVM	0.84±0.08	0.82±0.10	0.80±0.09	0.81±0.10	0.95±0.03
Naive Bayes+SMOTE	0.81±0.05	0.79±0.07	0.78±0.06	0.77±0.07	0.95±0.03
Naive Bayes	0.84±0.07	0.82±0.09	0.80±0.08	0.80±0.09	0.95±0.03
Logistic Regression+SMOTE	0.87±0.07	0.85±0.08	0.85±0.08	0.84±0.08	0.96±0.03
Logistic Regression	0.85±0.07	0.83±0.10	0.80±0.08	0.80±0.09	0.96±0.03
XGBoost+SMOTE	0.83±0.05	0.80±0.07	0.78±0.06	0.78±0.06	0.95±0.03
XGBoost	0.82±0.06	0.78±0.07	0.76±0.06	0.76±0.07	0.95±0.03
LightGBM+SMOTE	0.84±0.04	0.81±0.06	0.78±0.05	0.78±0.06	0.94±0.02
LightGBM	0.83±0.05	0.80±0.08	0.78±0.07	0.78±0.07	0.94±0.03
XGBoost+LightGBM+SMOTE	0.83±0.06	0.80±0.08	0.78±0.07	0.78±0.08	0.94±0.03
XGBoost+LGBM	0.83±0.05	0.79±0.07	0.78±0.06	0.78±0.06	0.94±0.03

As seen in the experimental results, the Logistic Regression classifier achieved the highest performance across all metrics. In addition, it was observed that applying SMOTE had a positive effect on all classifiers except Naive Bayes.

Table 10. Wilcoxon Signed-Rank Test Results for SMOTE vs. No-SMOTE Models

Classifier	Accuracy (p-value)	F1-score (p-value)	ROC-AUC (p-value)	Interpretation
Random Forest	0.063	0.063	0.492	Not Significant
SVM	1.000	0.578	0.734	Not Significant
Naive Bayes	0.031	0.055	0.275	Only Accuracy Significant
Logistic Regression	0.313	0.078	0.678	Not Significant
XGBoost	1.000	0.688	0.432	Not Significant
LightGBM	0.750	1.000	0.820	Not Significant
XGBoost+LGBM	0.250	0.844	0.492	Not Significant

The Wilcoxon signed-rank test was applied to compare the performance of models with and without SMOTE across cross-validation folds. The reported p-values indicate whether the differences in performance metrics are statistically significant at the 0.05 significance level.

The results of the Wilcoxon Signed-Rank test indicate that the application of SMOTE does not lead to statistically significant improvements across most classifiers and evaluation metrics ($p > 0.05$). Only the Naive Bayes model demonstrates a statistically significant improvement in

terms of accuracy ($p < 0.05$). These findings suggest that the impact of SMOTE on model performance is generally limited and model-dependent, providing measurable benefit only in specific cases rather than consistently enhancing performance across all classifiers.

4. Discussion

The results obtained from comparative analyses show that the Logistic Regression method demonstrated the highest performance in almost every metric, both in SMOTE and non-SMOTE settings. This indicates that it provides high performance in class separation.

Feature importance rankings in SMOTE and no-SMOTE configurations showed similar characteristics across all applied models. Features related to spinal alignment, particularly the dominance of spondylolisthesis degree, sacral tilt, and lumbar lordosis angle, are in good agreement with established clinical knowledge regarding spinal pathologies. This consistency demonstrates that the oversampling strategy improves prediction balance without altering the fundamental biomechanical relationships learned by the model.

The limited sample size of the dataset used is the most significant limitation of this study. Since the dataset belongs only to a specific group of patients, the model's generalizability to different populations remains limited. Furthermore, it is believed that including clinical, demographic, or laboratory data in addition to biomechanical properties would yield more decisive and reliable results. The study compared three classical machine learning algorithms. The methods applied could be supported by deep learning-based models or hybrid methods if the dataset could be further expanded. Finally, since the dataset used was not fully balanced, class distribution differences may have caused performance drops in some algorithms.

In comparative analyses, it is generally observed that class separation is successfully achieved in both SMOTE and non-SMOTE settings. The improvements observed in macro-average accuracy, precision, recall, and F1 scores after applying SMOTE demonstrate its effectiveness. SMOTE has been shown to improve macro-average sensitivity and F1 scores, particularly in underrepresented classes, by synthetically balancing minority classes; however, these improvements remained within the standard deviation range in some models.

Despite promising results, this study has several limitations that should be considered. First, the experiments were conducted on a limited dataset, which may limit the generalizability of the findings to broader clinical populations. Second, although SMOTE effectively addresses class imbalance, it relies on synthetic sample generation and may not fully capture the complex distribution of minority class samples in the real world.

Enhancing the dataset by increasing the number of real samples is a factor that could improve consistency.

The similarity in feature importance rankings across SMOTE and no-SMOTE configurations indicates the consistency of the Random Forest, XGBoost, and LightGBM models. Among the features related to spinal alignment, the degree of spondylolisthesis is particularly prominent in terms of feature importance. The consistent results obtained show that the model maintains a consistent prediction balance without altering the fundamental biomechanical relationships. This consistency is of significant clinical importance. The proposed approaches produce meaningful output not only in terms of classification performance but also in decision support systems for diagnosing spinal disorders.

Among all the machine learning models evaluated in this study, Logistic Regression consistently demonstrated the best overall performance across all evaluation metrics, particularly when combined with SMOTE, and achieved superior accuracy, recall, and F1 score values. Additionally, Random Forest and Support Vector Machine also demonstrated strong and stable performance with high ROC-AUC values and balanced classification results across classes. On the other hand, ensemble and boosting-based methods such as XGBoost, LightGBM, and their Soft Voting combination yielded competitive but not superior results compared to simpler models. These findings suggest that for this dataset, simpler and more interpretable models such as Logistic Regression may be preferred, especially when combined with class imbalance handling techniques like SMOTE.

The impact of SMOTE on ensemble-based models such as XGBoost and LightGBM has been widely reported in the literature as generally improving classification performance, particularly for imbalanced medical datasets. For instance, previous studies have shown that combining SMOTE with boosting-based and ensemble learning methods can enhance minority class detection and improve overall predictive performance.

Previous studies, such as the integration of SVM-SMOTE with ensemble learning for cerebrovascular disease prediction [38], have reported significant improvements in minority class detection and overall model performance, highlighting the effectiveness of combining data balancing techniques with ensemble methods. In the study Karamti et al. [39], SMOTE, when used in conjunction with ensemble learning and advanced preprocessing techniques, leads to improved performance. The highest accuracy of 99.9% was achieved when XGB, RF, and ETC were used together.

However, the findings of this study indicate a different trend. Although SMOTE provided slight improvements in certain metrics such as Precision and Recall, it did not result in a significant increase in overall performance for XGBoost, LightGBM, or their Soft Voting combination.

In particular, the macro-average ROC-AUC values remained largely unchanged after the application of SMOTE.

5. Conclusion

This study has demonstrated the effectiveness of machine learning methods in the early diagnosis and classification of orthopedic diseases. In the study, which was conducted using 7 different methods, accuracy, precision, recall, F1-Score, and ROC-AUC values were calculated. The results obtained indicate that biomechanical properties may be decisive in the diagnosis of orthopedic diseases. The most successful results were obtained using SMOTE with logistic regression, achieving an accuracy of 0.87, precision of 0.85, recall of 0.85, F1-Score of 0.84, and ROC-AUC of 0.96.

In addition to successful classification results, this study has some limitations that should be considered. The dataset is publicly available and limited. This may limit the generalizability of the findings to broader clinical populations. Furthermore, although SMOTE effectively addresses class imbalance, this method, which relies on synthetic sample generation, may not always fully capture the complex distribution of class samples in the real world.

It is anticipated that expanding the dataset, increasing the number of samples, and enriching the variety of biomechanical parameters will enhance the model's generalization capability. Furthermore, applying more complex deep learning models could further improve classification performance. It is recommended that future studies be expanded to include demographic factors such as different age groups and gender distributions.

Based on the data obtained from the evaluation of various criteria of orthopedic patients, this study can serve as a guide for disease detection and early diagnosis. It can also serve as a guide for similar studies to be conducted in the future. It is believed that more accurate results could be obtained if the content of the dataset could be expanded and the determining factors could be identified more clearly. The study makes an important contribution to the literature in terms of its concrete findings regarding disease detection.

To enhance the robustness and applicability of the proposed framework, future studies may evolve in various directions. The use of larger and more diverse datasets, including data from different populations and clinical settings, will improve the generalizability of the models. Incorporating clinical, demographic, and imaging data alongside biomechanical features could enable a more comprehensive representation of patient conditions and help achieve more consistent results. Furthermore, the application of advanced deep learning and hybrid model such as neural networks combined with ensemble learning strategies could further improve classification

performance. Another key aspect is the integration of explainable artificial intelligence (XAI) techniques to interpret model decisions and enhance clinical confidence.

Declaration of Ethical Standards

The authors declare that this study was conducted in accordance with accepted ethical standards in research and publication. All authors have contributed honestly to the work, appropriate citations have been provided, and the manuscript represents original research that has not been published or submitted elsewhere.

Credit Authorship Contribution Statement

All authors contributed to the conception and design of the study. Data analysis and model development were performed collaboratively. All authors participated in the interpretation of results, drafting, and critical revision of the manuscript, and approved the final version for submission.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding / Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The data used in this study are derived from the publicly available Biomechanical Features of Orthopedic Patients dataset.

References

- [1] A. M. Elshewey and A. M. Osman, "Orthopedic disease classification based on breadth-first search algorithm," *Scientific Reports*, vol. 14, no. 1, Art. no. 23368, 2024.
- [2] Y. Zhang, X. Liu, and H. Wang, "Spinal disease classification using SMOTE-RFE-XGBoost model," *PeerJ Computer Science*, vol. 9, e1280, 2023.
- [3] L. Frizziero et al., "New methodology for diagnosis of orthopedic diseases through additive manufacturing models," *Symmetry*, vol. 11, no. 4, Art. no. 542, 2019, doi: 10.3390/sym11040542.
- [4] S. Kıvrak, M. A. Aydın, and H. Polat, "Evaluation of machine learning models with SMOTE for imbalanced medical datasets," *Diagnostics*, vol. 14, no. 23, Art. no. 2634, 2024.
- [5] Z. Yang et al., "A hybrid machine learning approach using SMOTE and ensemble methods for healthcare prediction," *Scientific Reports*, vol. 15, Art. no. 92722, 2025.
- [6] Y. Gyasi-Agyei, "Comparative analysis of machine learning algorithms for medical diagnosis," *Results in Engineering*, vol. 21, Art. no. 101234, 2025.
- [7] S. Rezapour, M. H. Mahoor, and R. L. Figueroa, "Machine learning-based gait analysis for orthopedic disorder prediction using SMOTE," *arXiv preprint arXiv:2309.15990*, 2023.
- [8] H. B. Kibria and A. Matin, "The severity prediction of binary and multi-class cardiovascular disease—a machine learning-based fusion approach," *Comput. Biol. Chem.*, vol. 98, Art. no. 107672, 2022, doi: 10.1016/j.compbiolchem.2022.107672.

- [9] S. Li and X. Zhang, "Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm," *Neural Comput. Appl.*, vol. 32, pp. 1971–1979, 2020, doi: 10.1007/s00521-019-04378-4.
- [10] C. J. Ling et al., "Machine learning-based segmentation of images to diagnose orthopedic diseases and to guide orthopedic surgeries," *Soft Comput.*, 2023, doi: 10.1007/s00500-023-08503-3.
- [11] N. Rubaiyat et al., "Classification and prediction of orthopedic disease based on lumbar and pelvic state of patients," in *Proc. IEEE Int. Conf. Electr., Comput. Commun. Technol. (ICECCT)*, pp. 1–4, 2019, doi: 10.1109/ICECCT.2019.8869540.
- [12] N. Jahan et al., "Classification of orthopedic patients using supervised machine learning techniques," in *Intelligent Computing and Optimization*, Springer, pp. 659–669, 2021, doi: 10.1007/978-3-030-68154-8_61.
- [13] K. Hasan et al., "A machine learning approach on classifying orthopedic patients based on their biomechanical features," in *Proc. 7th Int. Conf. Informatics, Electron. Vis. (ICIEV) and 2nd Int. Conf. Imaging, Vis. Pattern Recognit. (icIVPR)*, pp. 289–294, 2018, doi: 10.1109/ICIEV.2018.8641004.
- [14] Q. Chen, Y. Zhang, M. Zhang, Z. Li, and J. Liu, "Application of machine learning algorithms to predict acute kidney injury in elderly orthopedic postoperative patients," *Clin. Interv. Aging*, pp. 317–330, 2023.
- [15] D. H. Mantzaris, G. C. Anastassopoulos, and D. K. Lymberopoulos, "Medical disease prediction using artificial neural networks," in *Proc. IEEE Int. Conf. Bioinformatics and Bioengineering*, pp. 1–6, 2008.
- [16] S. W. Chung et al., "Automated detection and classification of the proximal humerus fracture by using deep learning algorithm," *Acta Orthop.*, vol. 89, no. 4, pp. 468–473, 2018, doi: 10.1080/17453674.2018.1453714.
- [17] J. Olczak et al., "Ankle fracture classification using deep learning: Automating detailed AO/OTA 2018 malleolar fracture identification," *Acta Orthop.*, vol. 92, no. 1, pp. 102–108, 2020, doi: 10.1080/17453674.2020.1837420.
- [18] L. Cao, R. Li, D. Zhou, M. Zhao, and W. Huang, "Deep learning-based diagnosis and classification of femoral head necrosis," in *Proc. 5th Int. Conf. Artif. Intell. Ind. Technol. Appl. (AIITA)*, pp. 1225–1228, 2025.
- [19] A. M. Elshewey and A. M. Osman, "Orthopedic disease classification based on breadth-first search algorithm," *Sci. Rep.*, vol. 14, no. 1, Art. no. 23368, 2024.
- [20] UCI Machine Learning Repository, "Biomechanical features of orthopedic patients," 2023. [Online]. Available: Kaggle dataset.
- [21] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Inf. Sci.*, vol. 505, pp. 32–64, 2019.
- [22] P. Dutta, S. Paul, and M. Majumder, "An efficient SMOTE based machine learning classification for prediction and detection of PCOS," 2021.
- [23] Abacı, İ., & Yıldız, K. SMOTE vs. KNNOR: An evaluation of oversampling techniques in machine learning. *Gümüşhane Üniversitesi Fen Bilimleri Dergisi*, 13(3), 767-779, 2023.
- [24] Z. Sun et al., "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Syst. Appl.*, vol. 237, Art. no. 121549, 2024.
- [25] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Comput. Stat. Data Anal.*, vol. 52, no. 4, pp. 2249–2260, 2008, doi: 10.1016/j.csda.2007.08.015.
- [26] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [27] Y. Arora and S. K. Gupta, "Brain tumor classification using weighted least square twin support vector machine with fuzzy hyperplane," *Eng. Appl. Artif. Intell.*, vol. 138, Art. no. 109450, 2024.
- [28] S. Huang et al., "Applications of support vector machine learning in cancer genomics," *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.
- [29] A. Patle and D. S. Chouhan, "SVM kernel functions for classification," in *Proc. Int. Conf. Adv. Technol. Eng. (ICATE)*, pp. 1–9, 2013, doi: 10.1109/ICAdTE.2013.6524737.
- [30] O. N. Manjrekar and M. P. Dudukovic, "Identification of flow regime in a bubble column reactor with a combination of optical probe data and machine learning technique," *Chem. Eng. Sci.: X*, vol. 2, Art. no. 100023, 2019, doi: 10.1016/j.cesx.2019.100023.
- [31] A. G. Abdulameer, A. S. Hammood, F. M. Abdulwahed, and A. A. Ayyash, "Naïve Bayes algorithm for timely fault diagnosis in helical gear transmissions using vibration signal analysis," *Int. J. Interact. Des. Manuf.*, vol. 19, no. 5, pp. 3695–3706, 2025.
- [32] Ö. Tonkal and H. Polat, "Traffic classification and comparative analysis with machine learning algorithms in software defined networks," *Gazi Univ. J. Sci. C: Des. Technol.*, vol. 9, no. 1, pp. 71–83, 2021, doi: 10.29109/gujsc.869418.
- [33] B. Cömert, "Alın bölgesinden alınan elektrookülogram (EOG) işaretleri için ölçüm devresi tasarımı ve sınıflandırılması," M.S. thesis, Balıkesir Univ., Inst. Sci., Balıkesir, Türkiye, 2016.
- [34] H. Li et al., "Prediction of urban forest aboveground carbon using machine learning based on Landsat 8 and Sentinel-2," *Remote Sens.*, vol. 15, no. 1, Art. no. 284, 2023, doi: 10.3390/rs15010284.
- [35] A. Tasic et al., "Towards sustainable societies: Convolutional neural networks optimized by modified crayfish optimization algorithm aided by AdaBoost and XGBoost for waste classification tasks," *Appl. Soft Comput.*, vol. 175, Art. no. 113086, 2025.
- [36] S. Kakkar et al., "Enhancing energy efficiency and classification modeling through a combined approach of LightGBM and stratified k-fold cross-validation," *Electr. Power Compon. Syst.*, pp. 1–19, 2024.
- [37] S. Dörterler, "Hybridization of k-means and meta-heuristics algorithms for heart disease diagnosis," *New Trends Eng. Appl. Nat. Sci.*, p. 55, 2022.
- [38] Nithya, R., Kokilavani, T., & Beena, T. L. A. Balancing cerebrovascular disease data with integrated ensemble learning and SVM-SMOTE. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 13(1), 12, 2024.
- [39] Karamti, H., Alharthi, R., Anizi, A. A., Alhebshi, R. M., Eshmawi, A. A., Alsubai, S., & Umer, M. Improving prediction of cervical cancer using KNN imputed SMOTE features and multi-model ensemble learning approach. *Cancers*, 15(17), 4412, 2023.

*Research Article***Deep Learning-Based Lung Cancer Classification and Grad-Cam++ With Lime-Supported Explainability Analysis****Yasin DOGANSAH^{a,*} , Inayet Hakki CIZMECI^b , Murat KOKLU^c** ^a Department, of Computer Technologies, Afyon Vocational School, Afyon Kocatepe University, Afyonkarahisar, Türkiye^b Department of Computer Engineering, Afyon Kocatepe University, Afyonkarahisar, Türkiye.^c Department of Computer Engineering, Faculty of Technology, Selcuk University, Konya, Türkiye.

ARTICLE INFO

Article history:

Received 10 February 2026

Accepted 25 March 2026

*Keywords:*Explainable artificial
intelligence,Lung cancer,
Grad-CAM++,

LIME,

Medical image analysis.

ABSTRACT

This study aims to classify lung cancer using deep learning-based methods and to interpret the obtained model results using explainable artificial intelligence (XAI) approaches. The dataset used in this study consisted of 1933 chest computed tomography images, which were classified as normal, benign, and malignant. In the classification process, EfficientNet-B4, MobileNetV3-Large, and ResNet50 deep learning architectures were trained using a transfer learning approach, and their performance was evaluated using a 10-fold cross-validation method. The performance of the models was analyzed using accuracy, precision, recall, F1-score, and ROC-AUC metrics. According to the results obtained in the study, the MobileNetV3-Large model showed the highest overall performance with 97.05% accuracy and 99.70% ROC-AUC value. The EfficientNet-B4 and ResNet50 models also provided high and balanced performance values, achieving effective results in lung cancer classification. Grad-CAM++ and LIME methods were used to make model decisions more clinically reliable and interpretable. Grad-CAM++ analyses reveal that the models primarily focus on anatomically significant regions within the lung parenchyma during classification. LIME analyses, on the other hand, have enabled the superpixel-level explanation of local regions that contribute most to classroom decisions. Explainability maps obtained for normal, benign, and malignant classes showed that each class exhibited distinct spatial attention and contribution patterns. In conclusion, this study demonstrates that deep learning-based lung cancer classification offers a more reliable and transparent framework for clinical decision support systems, not only when evaluated with high-performance metrics but also when considered in conjunction with Grad-CAM++ and LIME-supported explainability analyses. The findings demonstrate that explainable artificial intelligence approaches play a significant role in improving model reliability in medical image analysis.

This is an open access article under the CC BY-SA 4.0 license.
(<https://creativecommons.org/licenses/by-sa/4.0/>)

1. Introduction

Globally, lung cancer is one of the leading causes of death and morbidity among all cancer types. This type of cancer is increasing every year, with an average of 2.5 million new cases diagnosed annually and causing an average of 2 million deaths per year [1, 2]. When lung cancer is diagnosed early and cancerous areas are removed, it becomes more difficult to treat in advanced stages [3]. In lung cancer, early diagnosis is as vital as

identifying the type of nodule detected early and removing it surgically [4]. Patients with advanced disease that cannot be surgically removed are treated with radiation therapy and chemotherapy. Despite this treatment, survival generally does not exceed 12 months. Due to the difficulty of treatment and rapid spread, the average 5-year survival rate for lung cancer is less than 20%. Therefore, early detection of lung nodules and complete removal of the mass before metastasis is critically important to increase survival rates [1, 3].

* Corresponding author. E-mail address: ydogansah@aku.edu.tr
DOI: 10.58190/ijamec.2026.164

Lung cancer has four different clinical stages. Stage I is the early diagnosis stage. Intervention at this stage increases the survival rate to 90%. Stage IV, the final stage, is the late diagnosis stage. At this stage, if treatment is delayed, the mortality rate reaches 90% [5]. Before initiating treatment, accurately determining the stage of the disease is critical for selecting the appropriate treatment method [6]. The first step in diagnosing lung cancer is a chest X-ray, which uses radiographs (chest X-rays) [7]. If a nodule is suspected after radiological evaluation of lung images, diagnosis is attempted using computed tomography (CT) or high resolution computed tomography (HRCT). In addition to these methods, other imaging techniques are also used for nodule characterization [8]. No matter how advanced these imaging methods are, lung anomalies can still be missed due to human error during the radiological evaluation process. Therefore, the use of computer assisted diagnostic systems aims to minimize human error. One of the primary goals of computer assisted diagnostic systems is to detect anomalies in images by using quantitative features extracted from those images. Additionally, by autonomously classifying the anomalies detected in the images as malignant or benign, it aims to support the clinical decision-making process in cases of differing opinions among experts [9-11]. In addition, computer aided systems such as deep learning approaches help analyze more data in a shorter time and make more accurate diagnoses compared to traditional anomaly detection methods [12]. Therefore, the success of deep learning in radiological image analysis and diagnoses is critical [13].

Explainable Artificial Intelligence (XAI), which offers transparency to users in AI-powered computerized diagnostic systems, refers to technical methods that enable users to understand the predictions produced by the models and the resulting outputs. [14]. Today, explainable artificial intelligence models are being used effectively in important fields such as finance, defense, and medicine. [15]. Artificial intelligence approaches, often described as "black boxes," can potentially lead to serious consequences, especially when used in vital fields such as medicine, due to the lack of transparency in decision-making processes and the inability to justify predictions. Therefore, the explainability of artificial intelligence models plays a significant role in medical applications. For healthcare professionals to trust artificial intelligence (AI) models during the diagnosis and treatment phase, the decision making capabilities of these AI models must be transparent, effective, and traceable [16].

This study aims to detect anomalies in the lungs using deep learning methods with computed tomography (CT) scans of the lungs. A deep learning model is used to first detect lung cancer and then determine whether the lung cancer is benign or malignant. It also aimed to ensure the

explainability of the findings by interpreting the results. This will increase the reliability of diagnoses and assist healthcare professionals in their diagnosis [17].

2. Literature Review

Lung cancer is one of the leading causes of cancer-related deaths worldwide, and early diagnosis directly impacts patient survival. Radiological imaging techniques are used as primary diagnostic tools in the detection of nodules. In recent years, the use of deep learning models in the analysis of radiological images has enabled high success rates in this field. However, the black box nature of these models creates problems of reliability and transparency, especially in clinical applications. In this context, explainable artificial intelligence (XAI) approaches are gaining importance in making model decisions interpretable.

In their 2017 study, Wang et al. developed a poorly controlled multi-label image classification and disease localization framework on the "ChestX-ray8" dataset, which contains 108,948 chest X-ray images from 32,717 unique patients. Natural Language Processing (NLP) techniques were used to automatically extract eight different disease labels from radiology reports: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax. In their study comparing different deep learning architectures, they achieved AUC values of 0.8141 for cardiomegaly, 0.7891 for pneumothorax, and 0.7362 for effusion using the ResNet-50 based model [18]. In 2017, Li and colleagues, using computed tomography (CT) images that provide detailed visualizations for more thorough examination and diagnosis of lung diseases, achieved an 83.03% success rate in predicting malignant lung cancer nodules with a hybrid model using 2620 CT images [19]. Mohandass and colleagues achieved 99.43% accuracy for cancer classification and 99.34% accuracy for non-cancerous classification with their AtCNN-DenseNet-201-TL-NBOA-CT hybrid model, developed using 888 CT scans and 1,186 lung nodules on the LUNA16 dataset [20].

In studies on lung cancer classification, Sivakumar et al. demonstrated high performance in 2024 by achieving 99.87% accuracy and 99.75% sensitivity with their ADBN + LightGBM (Adaptive Deep Belief Network) model applied to the LUNA16 and Kaggle DSB datasets [21]. Mohandass and colleagues achieved 99.43% accuracy in cancer classification and 99.34% accuracy in non-cancer classification with the AtCNN-DenseNet-201 TL-NBOA-CT optimized hybrid model developed using 888 CT scans and 1186 images from the LUNA16 dataset in 2024 [20]. In the field of segmentation and nodule detection, Budati and Karumuri achieved a 99.75% CPM in 2024 with their SbYSF (Sailfish + YOLO) based detection model developed using the LUNA16 dataset [22]. In terms of

segmentation, Chen et al. (2020) achieved a segmentation accuracy of over 99% with their proposed LDDNet model on the LIDC-IDRI dataset [23].

Studies on model efficiency and the use of different data types also hold a significant place in the literature. In their 2020 study, Masud and colleagues achieved 97.90% accuracy in terms of mobile device compatibility using a light CNN-based Light Deep Model on 1,279 images in the LIDC dataset [24, 25]. Focusing on multimodal imaging, Jose and her colleagues achieved 96.4% accuracy, 83.3% sensitivity, and 91.7% specificity with their DenseNet-based CNN model containing dilated convolutions, developed using 6,034 CT / X-ray images from the GitHub dataset [26]. In addition, Singh (2024), who investigated the effectiveness of hybrid architectures using attention mechanisms, obtained AUC values of 0.8556 for ChestX-ray14 and 0.904 for five pathologies in the CheXpert dataset with the SA-DenseNet121 model she developed on ChestX-ray14 and CheXpert datasets [27]. In their study conducted on a dataset containing 7560 chest X-ray (CXR) images, Veeramani et al. (2025) achieved 98% precision and 97% accuracy in diagnosing multi-class lung disease using an improved U-Net architecture for separating lung regions and an Inception-based XAI-TRANS transfer learning model for classification. [28]. Megat Ramli et al. (2025), supporting these high success rates obtained in individual studies at the general literature level, evaluated 34 studies on the detection of lung nodules in chest X-ray images and found that artificial intelligence algorithms showed significant potential in detecting small nodules, especially compared to radiologists (average AUC of 0.81), with sensitivity ranging from 56.4% to 95.7%, specificity ranging from 71.9% to 97.5%, and AUROC values ranging from 0.89 to 0.99 [29].

In studies to ensure the explainability of the models used, Naz et al. (2023) achieved a classification accuracy of 93% in the COVID-CT dataset and 97% in the COVID-Net dataset with a ResNet50-based transfer learning model they trained using COVID-CT and COVID-Net datasets to interpret lung diseases such as edema, tuberculosis, nodules, and pneumonia. They used the LIME technique to visualize the regions the model focused on, explain the classification results, and build clinical confidence [30]. Using a similar explainable artificial intelligence approach, Alqhatani et al. (2025) classified lung cancer stages (benign, malignant, and normal) on 1190 computed tomography (CT) images from the IQ-OTH/NCCD dataset using the EfficientNet-B0 architecture, achieving a high accuracy and precision rate of 99%, while generating visual heat maps using the Grad-CAM technique to make the model's decision processes transparent [31].

This study aims to analyze CT images of benign, malignant, and normal lung cancer using deep learning-based methods and to interpret the resulting classification decisions using explainable artificial intelligence

approaches. Although numerous classification studies on lung CT images exist in the literature, studies that explain the decision-making mechanism in a clinically transparent manner, as well as the model performance, are limited. In this context, the multi-class classification process, implemented using convolutional-based deep learning models, was supported by Grad-CAM++ and LIME methods, and the anatomical and pathological regions that the models focused on during decision-making were visually analyzed. The proposed approach contributes to the literature by not only improving classification success but also providing a explainable decision support framework that allows clinicians to evaluate the reliability of model outputs.

This article is structured as follows. The first section discusses the clinical significance of lung cancer, the necessity of early diagnosis, and the role of deep learning-based artificial intelligence approaches in this field. The second section summarizes current literature studies on lung cancer classification and explainable artificial intelligence. The third section details the dataset used, preprocessing steps, deep learning architectures, experimental environment, performance metrics, and explainability methods. In the fourth chapter, the experimental results obtained and the Grad-CAM++ and LIME-based explainability analyses are presented and discussed comparatively. The final section evaluates the overall results of the study and provides suggestions for future research.

3. Material and Methods

In this study, computed tomography (CT) images of lung cancer were obtained and analyzed using deep learning. In the first stage, the images were preprocessed to be compatible with deep learning models, and then multi-class classification was performed using EfficientNet-B4, MobileNetV3-Large, and ResNet50 architectures. The models' performance was analyzed using standard evaluation criteria such as accuracy, precision, sensitivity, F1-score, and ROC-AUC. The accuracy and generalizability of the proposed approach were evaluated using a 10-fold cross-validation method. Furthermore, to improve the interpretability of model predictions, Grad-CAM++ and LIME explanatory AI methods were applied, and regions contributing to classification decisions were visually analyzed.

3.1. Dataset Description

The dataset used in this study was created from chest computed tomography (CT) images collected from the National Chest Diseases Hospital (NIDCH) in Dhaka, Bangladesh, with the aim of supporting medical imaging research focused on the classification and early diagnosis of lung-related diseases through artificial intelligence and deep learning methods. This dataset, divided into three

diagnostic categories, contains a total of 1933 images: 1050 malignant, 467 benign, and 416 normal cases [32]. This dataset is important for training, validating, and comparing AI-based models developed for autonomous diagnosis to detect respiratory diseases and lung anomalies. In healthcare settings, the software contributes to the development of clinical decision support systems.

3.2. Dataset Preprocessing

In this study, a dataset consisting of lung computed tomography scans representing different pathological conditions was first obtained. The dataset, after being checked, was uploaded to Google Drive and transferred to the Google Colab platform. Next, the images were processed through standard data preprocessing steps to meet the input requirements of deep learning models, and the dataset was divided into training and validation subsets. Lung CT images were converted to RGB format, rescaled to fixed sizes depending on the architecture used (380×380 for EfficientNet-B4, 224×224 for MobileNetV3-Large and ResNet50), and converted to tensor format. To ensure compatibility with pre-trained ImageNet weights, the images were normalized using the mean and standard deviation values from the ImageNet dataset.

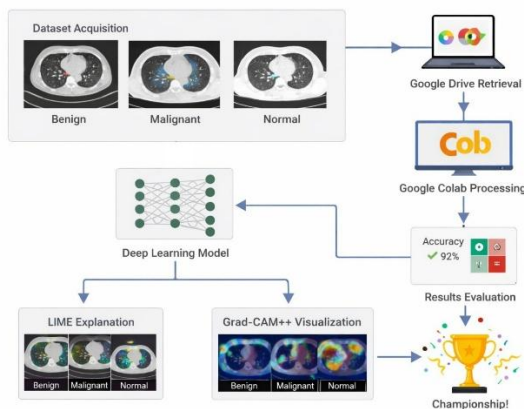


Figure 1. Data preprocessing

Figure 1 presents a deep learning and explainable artificial intelligence workflow (AI-generated) for autonomously distinguishing benign, malignant, and normal classes from lung CT images.

In the classification phase, three different deep learning-based convolutional neural network architectures—EfficientNet-B4, MobileNetV3-Large, and ResNet50—were used, and the models were trained using transfer learning. These architectures are optimized to autonomously learn distinctive structural and textural features in lung computed tomography images. Following the training process, the performance of the models was analyzed using evaluation metrics such as accuracy, precision, sensitivity, F1-score, and ROC-AUC. To improve the interpretability of model decisions, explainable artificial intelligence approaches have been

utilized. Using the Grad-CAM++ method, a demonstrable artificial intelligence approach, critical regions that the models focused on during classification were visualized through heat maps. With the LIME method, images were analyzed on a superpixel basis to identify local regions that contributed most to classification decisions. These visualizations have allowed for a more transparent and meaningful clinical evaluation of the models' decision-making mechanisms. In conclusion, these proposed approaches ensure not only high classification performance but also that the model outputs are explainable and reliable.

Table 1. Model Structure and Training Parameters

Parameter	EfficientNet-B4	MobileNetV3-Large	ResNet50
Model Architecture	EfficientNet-B4	MobileNetV3-Large	ResNet50
Pretrained Weights	ImageNet (Imagenet1k_V1)	ImageNet (Imagenet1k_V1)	ImageNet (pretrained=True)
Input Image Resolution	380 × 380	224 × 224	224 × 224
Number of Classes	Dataset-derived (ImageFolder)	Dataset-derived (ImageFolder)	3 (Benign, Malignant, Normal)
Loss Function	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss
Optimizer	AdamW	AdamW	Adam
Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Batch Size	16	16	16
Number of Epochs	3	5	5
Cross-Validation Strategy	10-Fold	10-Fold	10-Fold Stratified
Evaluation Metrics	Accuracy, F1-score, Precision, ROC-AUC Recall,	Accuracy, F1-score, Precision, ROC-AUC Recall,	Accuracy, F1-score, Precision, ROC-AUC Recall,
ROC-AUC Calculation	OvR (multi-class)	OvR (multi-class)	OvR (multi-class)

Table 1 presents the model structure and training parameters of the EfficientNet-B4, MobileNetV3-Large, and ResNet-50 deep learning architectures used for multi-class classification of computed tomography images of lung cancer. Each model is configured taking into account the input image resolution specific to its architectural structure. Computed tomography images have been rescaled to match the input dimensions expected by the relevant networks. The models were trained on the ImageNet dataset using pre-trained weights, and their performance was evaluated using standard grading measures such as accuracy, precision, sensitivity, F1-score, and ROC-AUC.

3.3. Experimental Environment and Application Details for Lung Cancer CT Imaging

Analysis of lung computed tomography images using deep learning models was performed using data access provided via Google Drive and in the Google Colab

environment with A100 GPU infrastructure.

Table 2. Computing Environment and Software Structure

Environment Details	Specifications (According to the Code)
Platform	Google Colab
Storage	Google Drive (mounted dataset directory)
Backend Framework	PyTorch
Programming Language	Python 3.10
GPU	NVIDIA A100 (CUDA-enabled, Google Colab)
Device Selection	Automatic (CUDA if available, else CPU)
Operating System	Linux (Google Colab environment)
Deep Learning Libraries	Torch, TorchVision
Data Processing	TorchVision Transforms
Visualization Library	Matplotlib
Evaluation & Metrics	Scikit-learn
Statistical Analysis	SciPy (paired t-test)
Cross-Validation Method	10-Fold Cross-Validation

Table 2 shows that the analysis studies performed using CT images of lung cancer were carried out in the Google Colab environment using Python 3.10 and a PyTorch-based deep learning infrastructure. TorchVision was used for data preprocessing, Matplotlib for visualization, Scikit-learn for performance evaluation, and SciPy for statistical analysis. All experiments were conducted on a CUDA-enabled NVIDIA A100 GPU using a 10-fold K-Fold cross-validation approach.

3.4. Performance Metrics

Performance metrics are criteria used to measure the accuracy and effectiveness of the artificial intelligence model being used, and they vary depending on different machine learning tasks such as classification, regression, and clustering [33]. The success of artificial intelligence models used in lung cancer diagnosis and medical imaging analysis is typically evaluated using statistical metrics calculated from an error matrix [31]. This matrix consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values [21].

3.4.1. Accuracy

Accuracy is a key evaluation metric that represents the percentage of correct predictions among all predictions made by the model and summarizes its overall performance (Veeramani et al., 2025). In medical diagnoses, it is calculated by dividing the number of correctly classified cases by the total number of cases [31]. The accuracy formula is presented in Equation (1)[34].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

3.4.2. Precision

Precision identifies the percentage of true positives among all cases that an AI model classifies as positive,

reflecting the model's ability to make accurate predictions [22]. This metric is used to evaluate the quality and reliability of the model, especially in situations where keeping the false positive rate low and preventing misdiagnosis is important [21, 22]. The equation of certainty is given in Equation (2) [35].

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3.4.3. Recall

Recall demonstrates how well a model can capture truly positive examples (Tian et al., 2024). Since it is vital not to miss a disease in medical diagnosis, the recall value is one of the most important parameters determining the sensitivity of the model, and the sensitivity equation is given in equation (3) [28, 36, 37].

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

3.4.4. F1-Score

The F1-Score is the harmonic mean of the precision and recall metrics, and the F1-Score equation is given in equation (4) [38]. It provides a more reliable performance measurement than accuracy rate alone, especially when the classes in the dataset are unevenly distributed [13, 39].

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

3.4.5. Specificity

Specificity is the model's ability to accurately identify true negative cases (Sivakumar et al., 2024). It is used to exclude the possibility of a healthy individual being mistakenly classified as ill [26]. The specificity equation is given in equation (5) [40].

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

3.5. Cross Validation

Cross-validation improves the generalization capabilities of machine learning models by evaluating their performance. It is a resampling method that allows for the acquisition of statistically reliable results [41]. This approach works on the principle of dividing the dataset into K equal-sized subsets called folds [42]. Throughout the process, in each iteration, one of these subsets is reserved as the validation set, while the remaining K-1 subsets are used for training the model [43]. This cyclical structure increases data utilization efficiency by ensuring that each instance in the dataset is included in both the training and testing phases throughout the process [9]. Furthermore, cross-validation allows the average of results obtained from multiple evaluations to minimize biases that may arise from irregularities in data distribution, thus ensuring that the model exhibits more stable and reliable performance [13].

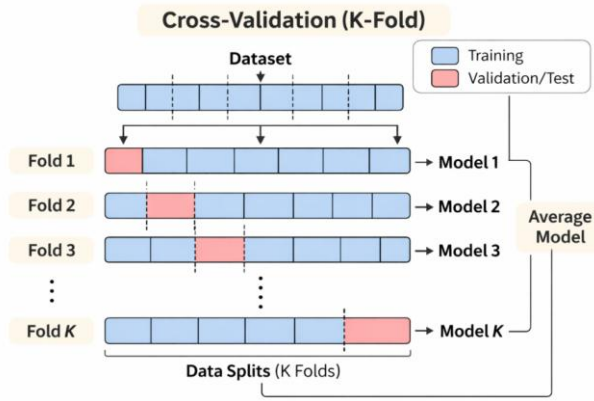


Figure 2. Schematic representation of the K-fold cross-validation process on lung CT images.

Figure 2 shows the k-fold cross-validation process. The dataset is divided into multiple folds. At each step, one fold is used for validation or testing, while the remaining folds are used for training. At each step, one fold is used for validation or testing, while the remaining folds are used for training. The images selected for validation or testing are changed in each iteration, ensuring the model is evaluated fairly and reliably across the entire dataset [44].

3.6. Explainable Artificial Intelligence (XAI)

Explainable AI is a method that makes the closed, often described as black box, nature of complex deep learning models transparent and understandable to humans [28]. In life-threatening medical diagnostic processes, such as lung cancer, a model that only produces predictions is not sufficient. Allowing physicians to see the steps involved in reaching this result increases both trust in the system and clinical accuracy [30]. XAI techniques aim to transform artificial intelligence from a decision-maker into a system that can present its reasoning and support it with visual evidence [28]. This approach helps artificial intelligence gain wider acceptance in healthcare by solving critical issues such as reliability, transparency, and algorithmic bias detection in diagnostic processes [28, 30].

3.6.1. LIME

Local Interpretable Model-Independent Explanations (LIME) is a model-independent method that explains the prediction of any AI model on a given sample at the local level. [31, 45]. This method divides a given image into smaller pieces that share similar visual characteristics and are called superpixels. Then, by randomly swapping these components, they analyze the model's responses [30, 31]. The LIME method creates a simple surrogate model that represents a complex pattern at the local level, highlighting the regions that contribute most to the prediction result through colored patches [30, 45]. In conclusion, LIME highlights key features of the model's decision-making process by directing the physician's focus to areas of pathological significance [30].

The mathematical optimization formula that forms the basis of the LIME architecture is given in equation (6) [30,

45]

$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (6)$$

Definitions of the components included in the formula;
 f : The actual black box that needs to be explained represents the classifier or model.

$g \in G$: It creates a simple surrogate model, such as linear regression, selected from the class of interpretable models (G), which represents the original model at the local level.

$L(f, g, \pi_x)$: The loss function measures the difference between the original model f and the surrogate model g , indicating how faithful the surrogate model is to the original model.

π_x : It is a weighting mechanism that determines the neighborhood and proximity of a given sample.

$\Omega(g)$: This term measures the complexity of a proxy model (e.g., number of rules or density of weights) so that the explanation remains understandable to humans [30, 45].

3.6.2. Gradient Weighted Class Activation Mapping (Grad-CAM)

Grad-CAM is an explainable artificial intelligence technique used to visualize complex decision-making processes in convolutional neural networks (CNNs) and reveal which regions directly influence model predictions [41]. In medical imaging, this method allows the model to move from being a black box to a transparent model by identifying pathological areas, such as nodules or lesions in the lung, that give rise to its prediction [12, 31].

The Grad-CAM method combines feature maps with gradient signals by retrospectively tracing the model's decision-making process. This approach contributes to evaluating the reliability of the diagnosis by visually demonstrating that the network focuses on medically critical regions. The Grad-CAM method preserves the spatial features in the final convolutional layers of the model, revealing which pixels contribute to the prediction score via a transparent heat map [27, 30]. The integration of these visualization techniques hierarchically explains the black-box nature of AI, thereby increasing healthcare professionals' clinical confidence in the model [12]. Heat maps generated by Grad-CAM help to verify the reliability of diagnosis by visually proving whether artificial intelligence is focusing on medically significant tumor or lesion areas [28, 31].

The Grad-CAM process mathematically takes place in two fundamental stages:

Calculation of Neuron Significance Weights (α_k^c): First, the gradient of y^c , which is the score of a class such as the target, is taken with respect to the activations of feature map number k (A^k) in the last convolutional layer of the model [28]. These gradients are subjected to a global average pooling process to obtain a single significance value for each feature map (Alqhatani et al., 2025).

The weight formula is given in equation (7).

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (7)$$

In equation (7), A_{ij}^k represents the activation map of the k convolution channel, y^c represents the model output of the c class, Z represents the total number of pixels in the activation map, and α_k^c represents the class specific weight of the relevant channel [46].

Creating a Heat Map: Secondly, the calculated weights are multiplied by the corresponding feature maps and then summed. Based on the obtained result, the ReLU (Linear Smoothing Unit) activation function is applied (Veeramani et al., 2025). The ReLU function ensures that spatial regions that positively contribute to the prediction of the target class are preserved, preventing the inclusion of negative activations that reduce the class score in the visualization (Alqhatani et al., 2025).

The heat map formula is given in equation (8) [47].

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \quad (8)$$

The heat maps generated through Equation (8) are overlaid on the original input image to visualize the critical regions that the model gives weight to in the decision mechanism [12]. In clinical practice, this visual evidence allows physicians to verify that artificial intelligence is focusing on medically relevant areas and to identify errors [9].

3.6.3. Grad-CAM++

Grad-CAM++ is an enhanced extension of the standard Grad-CAM method, specifically designed to more precisely determine the position of objects within an image and to identify situations where multiple disease markers are present [9, 30]. This enhanced version uses higher order combinations of gradients to more clearly and comprehensively visualize complex pathological features in medical images, such as ground glass opacities. Compared to the standard method, it offers physicians a more detailed and explanatory visual analysis by marking the regions contributing to the model's prediction with more accurate coverage [30]. Especially in cases where multiple lesions are present in a single image, it provides superior interpretability in clinical decision support processes by better differentiating the contribution of each object [9].

3.7. Deep Learning Algorithms

Deep learning algorithms with different architectural features have been used in the multi-class classification of CT images of lung cancer. The selected models were chosen for their ability to automatically learn distinctive features from images, effectively represent structural and textural information at different scales, and allow for comparative evaluation of their classification performance. Each architecture was structured according

to its own design principles and trained on the same dataset, and its performance was analyzed using standard evaluation criteria. In this context, EfficientNet, which stands out with its combined scaling approach, ResNet50, which aims to reduce the gradient loss problem thanks to its deep network structures, and MobileNetV3 Large, which provides computational efficiency with its lightweight structure, have been examined.

3.7.1. EfficientNet

The EfficientNet architecture is built on a unique composite scaling strategy that scales the network's depth, width, and resolution simultaneously and in a balanced manner [31]. EfficientNet-B4, as a member of this model family ranging from B0 to B7, is a mid-to-high-end model that aims to deliver higher performance while optimizing the increase in computing cost [25, 31].

3.7.2. Residual Network 50 (ResNet50)

ResNet50 is an advanced convolutional neural network architecture with a depth of 50 layers, developed in 2015 by Kaiming He and colleagues at Microsoft Research Asia. This network structure consists of 48 convolution layers, one maximization pooling layer, and one averaging pooling layer [28]. The most distinctive feature of the ResNet50 architecture, hopping connections, allows input to be passed directly to subsequent layers by skipping some layers. This solves the gradient disappearance problem encountered in deep networks [28, 48]. ResNet50 has over 23 million trainable parameters and is commonly used for transfer learning with pre-trained weights on massive datasets like ImageNet [12, 28]. Thanks to its identity matching capability, the model reduces the risk of overfitting the dataset by bypassing layers it deems unnecessary during training [48]. In medical imaging studies, it is frequently used as a teaching model due to its high accuracy and rapid convergence capabilities, particularly in complex tasks such as the classification of lung cancer subtypes and pathological image analysis [12, 13].

3.7.3. MobileNetV3-Large

The MobileNetV3-Large architecture is built upon depth-based decomposable convolutions and linear bottleneck inverted residual structures, which formed the basis of its predecessors, the V1 and V2 models [42]. This architecture is structured to include more layers and parameters compared to the lighter, smaller version of the family, specifically to achieve high accuracy in complex visual recognition tasks. The NetAdapt algorithms used in the design process allow each layer of the network to be optimized to operate most efficiently on the hardware. This effectively reduces calculation costs and latency. Thanks to its lightweight and modular design, MobileNetV3-Large can offer high performance with low resource consumption in a wide range of applications, from medical diagnostics to industrial quality control

processes [41]. The architectural advantages offered by MobileNetV3-Large provide an effective solution, especially in environments with limited computing resources such as mobile devices, thanks to the balanced structure it establishes between accuracy and latency [13]. The depth-based decomposable convolution structure reduces mesh parameters, increasing speed in computationally intensive fields such as medical image analysis and enabling more efficient processing of complex data [13, 20]. In critical clinical tasks such as lung cancer diagnosis, the use of such optimized models accelerates diagnostic processes, provides reliable opinions for radiologists, and strengthens decision support mechanisms [13].

4. Result and Discussion

Computed tomography images of lung cancer were classified into benign, malignant, and normal categories, and EfficientNet-B4, ResNet50, and MobileNetV3-Large models were applied to these images. The experimental results obtained were compared using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics in the multi-class classification problem and are presented in Table 3.

Table 3 Average Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
MobileNetV3-Large	0.9705 ± 0.0179	0.9688 ± 0.0206	0.9744 ± 0.0138	0.9706 ± 0.0165	0.9970 ± 0.0027
EfficientNet-B4	0.9617 ± 0.0161	0.9592 ± 0.0154	0.9624 ± 0.0266	0.9599 ± 0.0209	0.9704 ± 0.0193
ResNet50	0.9555 ± 0.0174	0.9577 ± 0.0162	0.9555 ± 0.0174	0.9555 ± 0.0172	0.9924 ± 0.0045

The MobileNetV3-Large model achieved the highest overall classification success among all models examined, with an accuracy rate of $97.05 \pm 1.79\%$. The fact that the model's Precision ($96.88\% \pm 2.06\%$), Recall ($97.44\% \pm 1.38\%$), and F1-Score ($97.06\% \pm 1.65\%$) values are quite close to each other indicates a balanced distribution between false positive and false negative predictions and that the model has a consistent decision-making mechanism. Furthermore, the $99.70\% \pm 0.27$ ROC-AUC value reveals that the MobileNetV3-Large model has a very high inter-class discrimination capability.

The EfficientNet-B4 model achieved a performance close to MobileNetV3-Large with an accuracy rate of $96.17\% \pm 1.61$. The fact that the Precision ($95.92\% \pm 1.54\%$), Recall ($96.24\% \pm 2.66\%$), and F1-Score ($95.99\% \pm 2.09\%$) metrics are close to each other indicates that the model exhibits balanced and stable classification behavior. However, the ROC-AUC value of $97.04\% \pm 1.93$ indicates that the EfficientNet-B4 model successfully distinguishes between classes.

The ResNet50 model, with an accuracy rate of $95.55 \pm$

1.74% , offered a relatively lower overall accuracy compared to the other two models. The fact that the Precision ($95.77\% \pm 1.62\%$), Recall ($95.55\% \pm 1.74\%$), and F1-Score ($95.55\% \pm 1.72\%$) values are very close to each other indicates that the model exhibits stable classification performance. However, the ROC-AUC value ($99.24\% \pm 0.45\%$) indicates that the ResNet50 model demonstrates strong performance in terms of inter-class discrimination.

Overall, the MobileNetV3-Large model stands out in terms of overall performance, achieving both the highest accuracy and the highest ROC-AUC values. While the EfficientNet-B4 model offers balanced and high performance, the ResNet50 model provides a strong distinction, especially in the ROC-AUC metric. These results demonstrate that lightweight and efficient convolutional architectures (MobileNet and EfficientNet) offer effective solutions in terms of both accuracy and stability on the relevant dataset.

4.1. Analysis of Lung CT Images Using Grad-CAM++ Method

Explainable artificial intelligence approaches have been applied to make the predictions obtained from the deep learning-based classification model more reliable and interpretable from a clinical perspective. The Grad-CAM++ method, which generates class-separated spatial attention maps to reveal which regions of the image the model focuses on when making decisions, was used. In addition, LIME analysis was performed by dividing the image into superpixel regions, locally explaining the contribution of each region to the classification result. Thus, for all three classes, Grad-CAM++ shows both where the model concentrates and how each region influences the class decision. With LIME, the model's decision-making mechanism is supported by visual evidence through comparative evaluation.

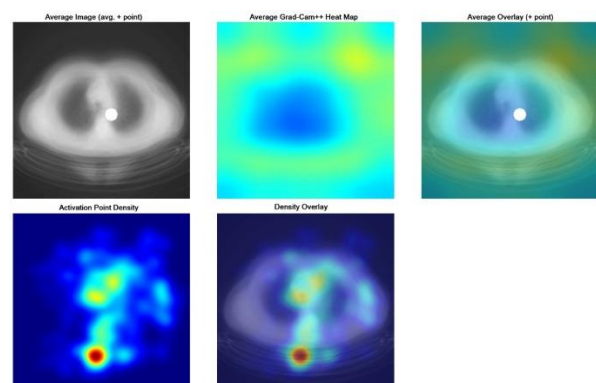


Figure 3. Average Grad-CAM++ activation and density analyses belonging to the normal class.

Figure 3 shows the average Grad-CAM++ activation maps generated from all lung CT images belonging to the Normal class (images without disease symptoms). The images obtained revealed low-intensity and diffuse

activation patterns throughout the lung parenchyma. The model does not appear to focus strongly on a specific anatomical region; the classification decision is based on the absence of pathological structural cues. This indicates that the model does not produce abnormal focuses or random attention areas when making decisions in normal cases.

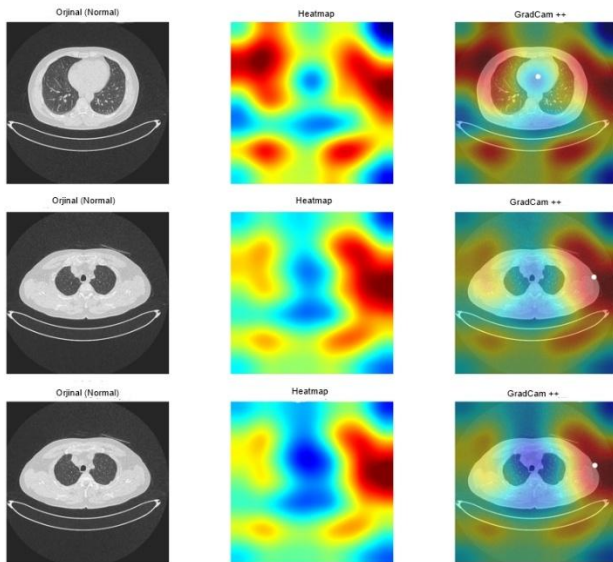


Figure 4. Example-based Grad-CAM++ visualizations belonging to the standard class.

In Figure 4, it is observed that the average activation point (Average activation point: 124.16, 117.36) is concentrated in the coordinates of the model input scale. The fact that this point corresponds to the central regions of the lung area indicates that the model has developed an anatomically consistent and non-random attention mechanism for the normal class. In the images, the white dot represents the average activation center calculated from the Grad-CAM++ heatmap. The model shows the centroid of the region that contributes most spatially to the classification decision.

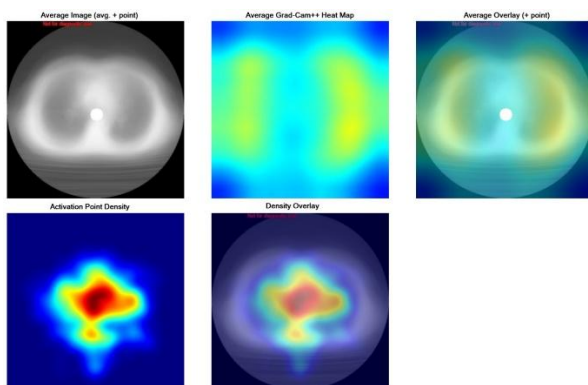


Figure 5. Average Grad-CAM++ activation and density analyses belonging to the malignant class.

Figure 5 presents the average images, Grad-CAM++ heat maps, and activation intensity analyses obtained from the analysis of malignant class lung CT images using a deep learning model. The average CT images and

corresponding Grad-CAM heat maps presented in the upper panel demonstrate that the model focuses its decision-making attention on distinct, asymmetric, and localized pathological regions within the lung parenchyma. The activation point density and density overlay visualizations in the lower panel clearly show that the regions contributing most to the classification decision emerge as spatially consistent hotspots. This indicates that, in contrast to the homogeneous and symmetrical attention distribution observed in the normal class, the structural distortions and lesion-like morphological abnormalities specific to the malignant class are effectively distinguished by the model. The analysis reveals that by highlighting the critical regions focused on by the deep learning-based diagnostic model, both the model's transparency is increased and it is possible to interpret the locations where pathological findings are most frequently concentrated in a clinically meaningful way.

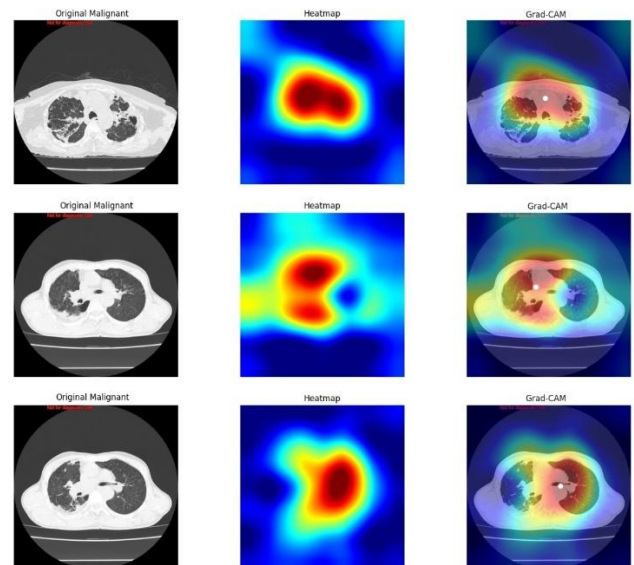


Figure 6. Example-based Grad-CAM++ visualizations of the Malignant class.

Figure 6 shows three different axial CT images of malignant lung cancer. The Grad-CAM results obtained from these images significantly reflect the model's decision-making mechanism. In the first example, the model's production of high red activation in the middle and lower regions of the lung parenchyma suggests that it is focusing on potentially malignant areas. The blue tones observed in extrapulmonary structures indicate that these regions have low importance in the decision-making process. In the second example, while significant activations were observed in pathological regions located in both lungs, the presence of low-level activations in peripheral and non-tumor-related areas indicates that the model exhibits careful and selective focusing. In the third example, the activations are predominantly concentrated within the lung parenchyma, with weaker activations on the opposite side. These findings demonstrate that the

model is able to accurately distinguish anatomical regions associated with malignancy and exhibits generally consistent classification behavior.

Grad-CAM++ heat maps show that the model primarily focuses on regions corresponding to abnormal tissue patterns and potential lung nodules within the lung parenchyma. In malignant cases, highlighted areas generally overlap with regions where nodular structures are visually prominent, i.e., regions with varying densities. These findings are consistent with pathological patterns reported in the healthcare field for lung anomaly detection. This observation demonstrates that the model's decision-making process relies on clinically significant image features (density differences) rather than irrelevant regions, thus supporting the reliability of the proposed explainable AI framework.

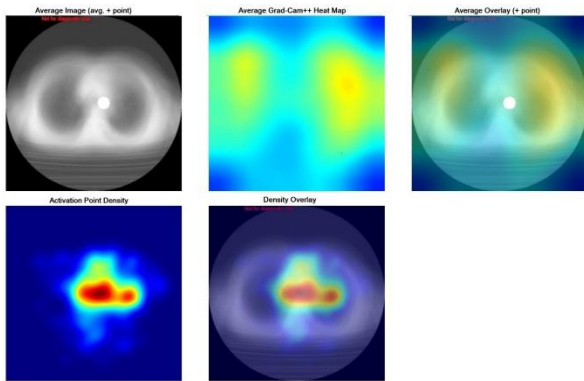


Figure 7. Average Grad-CAM activation and intensity analyses belonging to the benign class.

Figure 7 shows the average image obtained from benign class lung CT sections, the corresponding average Grad-CAM++ heat map, overlay, and activation point density. The average Grad-CAM++ map reveals that, unlike the sharp, singular hotspot structure seen in the malignant grade, the model's attention is focused on a broader region extending into the central parenchymal areas of the lungs. Particularly in the overlay panel, it is observed that the activation largely corresponds to intrapulmonary structures, while there is no significant concentration in extrapulmonary regions such as the chest wall and background. The fact that the hot regions in the activation point density map exhibit a smoother, more balanced profile suggests that the model's decision in the benign class is supported by the overall textural arrangement of the parenchyma and widespread structural cues, rather than by a single mass-like focus. The calculated average activation point (mean activation point: 123.58, 110.73) quantitatively confirms that this focus corresponds to the middle sections of the lungs. Thus, it appears that the benign class is represented by a more stable and diffused attention distribution compared to the malignant class.

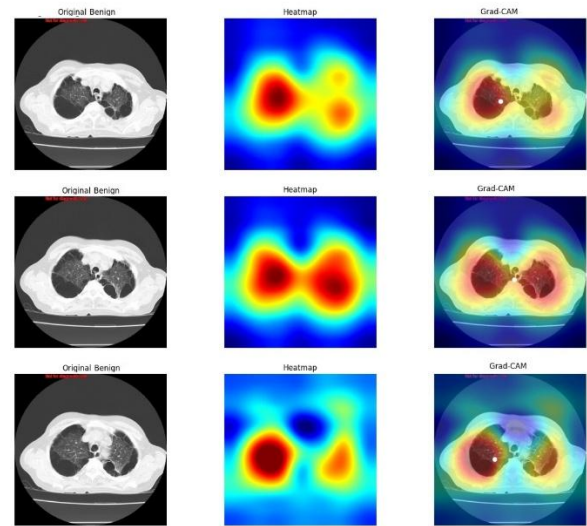


Figure 8. Example-based Grad-CAM visualizations belonging to the benign class.

Grad-CAM++ based attention maps of three benign (non-cancerous) lung CT samples presented in Figure 8 show that the model's classification decision is predominantly based on structures within the lung parenchyma. In the first example, although the dominant activation is more pronounced on one side, a secondary contribution is also observed in the opposite lung, suggesting that the decision is formed by a combination of features distributed throughout the parenchyma rather than a single, sharp focus. In the second example, the appearance of activation areas of similar size and intensity in both lungs suggests the adoption of a more balanced spatial attention pattern characteristic of the benign class. In the third example, although activation is more dominant in one lobe, it does not completely subside in the opposite lobe but persists at a moderate level, supporting the idea that the model makes its decision by evaluating bilateral tissue patterns together, rather than locking onto a local aggressive focus. Overall, the concentration of the highest contribution in the parenchyma rather than in extrapulmonary structures indicates that the obtained Grad-CAM++ samples are anatomically consistent and strengthens the interpretability of the benign class.

4.2. Analysis of Lung CT Images Using the LIME Method

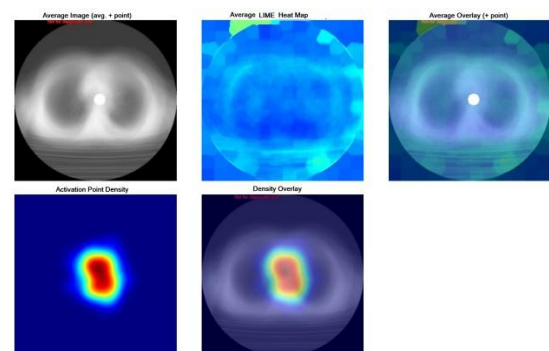


Figure 9. Average LIME activation and intensity analyses belonging to the Benign class.

Figure 9 shows the results of the average activation analysis obtained using the LIME method for the benign lung class. The average CT image in the upper left represents the spatial common structure of benign samples in the dataset, while the marked center point reveals the location most frequently referenced by the model in the decision-making process. The average LIME shows that the additive regions in the heat map exhibit a superpixel-based distribution with smooth transitions, without forming sharp boundaries. This suggests that the model, when defining the benign class, relies on structural patterns formed by neighboring tissues rather than a single dominant lesion. In the overlay visualization, it is noteworthy that these contribution areas remain largely within the lung parenchyma and extend only minimally into surrounding anatomical structures. The activation point density map reveals that the high contribution is concentrated in a narrow and compact region, but exhibits a profile different from the aggressive and widely distributed foci observed in malignant classes. Finally, the density overlay panel shows that the model uses a central and repetitive spatial attention mechanism when making decisions for the benign class, but that this attention is widely and evenly distributed. These results indicate that LIME-based average analyses offer a more stable model and structurally-driven explainability in benign samples.

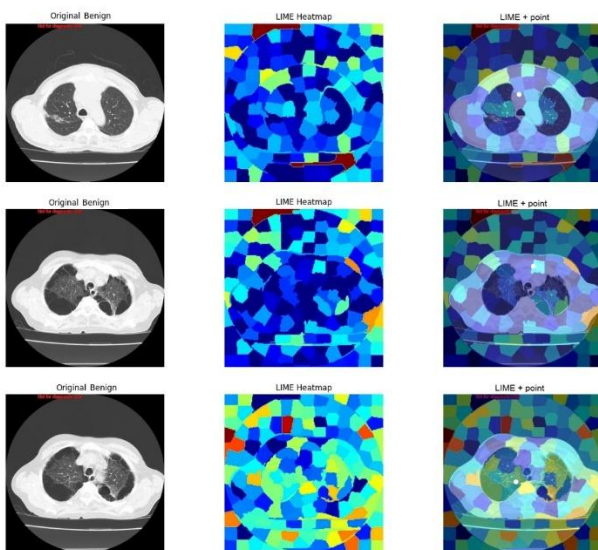


Figure 10. Benign class-specific example-based LIME visualizations

Figure 10 shows an example of a benign lung computed tomography image, illustrating how the LIME-based explainability approach shapes model decisions. In the first example, the general structure of the lung parenchyma is preserved in the original CT scan, and no significant aggressive lesion pattern is observed. In contrast, the LIME heatmap shows the contributions spread across numerous small superpixels, with low to medium dispersion. The LIME visualization shows that the model's decision center is located in the central region of the lungs, but the decision is not based on a single dominant region.

In the second example, while increased heterogeneity in the parenchyma is noticeable in the original image, it is striking that the contributions in the LIME heat map are mostly limited to intrapulmonary tissues, and peripheral areas are represented with low significance. This suggests that the model bases its definition of the benign class on structural tissue arrangement. The LIME output shows that the insertion center is located within the parenchyma, supporting anatomical consistency. In the third example, although there are more pronounced density differences in the original section, the LIME heat map shows that the contributions are spread over a wide area and do not form a sharp focus. In the images in the third column, the location of the maximum contribution, marked with a dot, reveals that the model based its decision on the combined effect of environmental superpixel groups rather than a specific local structure. Overall, these three examples demonstrate that the LIME method explains the model's decisions for the benign class in a piecemeal, distributed, and structurally-based manner, deviating significantly from the dense and localized contribution patterns expected in the malignant classes.

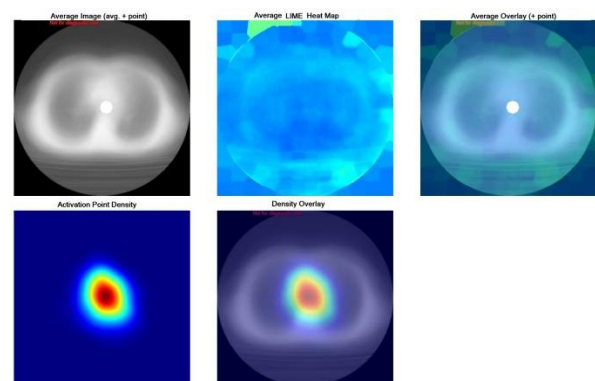


Figure 11. Average LIME activation and intensity analyses belonging to the Malign class.

Figure 11 shows the average activation analysis obtained using the LIME method for the malignant class. The average CT image in the upper left panel represents the common spatial appearance of malignant specimens, while the marked point indicates the most frequently observed decision center within each class. In the average LIME heat map (upper center), the contribution regions are distributed with a superpixel-based and fragmented structure, but the most prominent contribution is concentrated in regions near the center of the image. This suggests that the model's decision is based on the combined effect of multiple neighboring structural components, rather than a single small region. The overlay panel in the upper right visually demonstrates how the model's attention is drawn to the central structures within the image by presenting these contribution areas superimposed on the anatomical structure. The activation

point density map in the lower left shows that activations are clustered distinctly in a single core region, and the calculated average activation point (average activation points: 116.52, 111.05) quantitatively confirms the center of this clustering. Finally, the density overlay panel (lower middle) clarifies where the high contribution is repeated on the image, demonstrating that the model has developed a more stable spatial focus in the lung malignant nodule class.

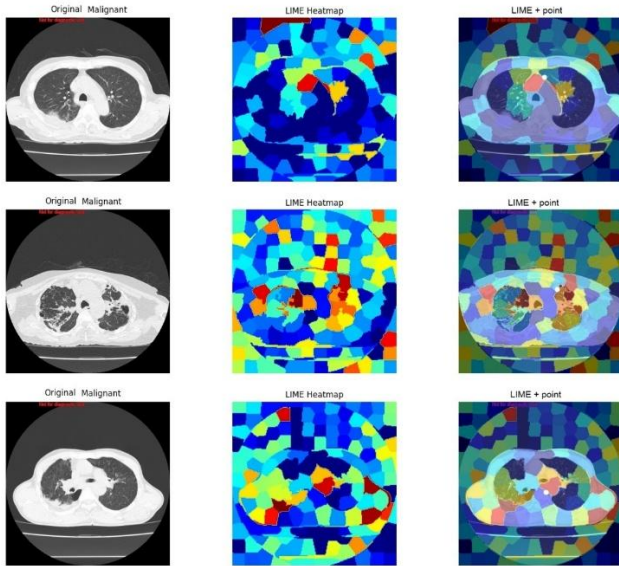


Figure 12. Example-based LIME visualizations of the malignant class.

Figure 12 illustrates how the model justifies its decision in the malignant classification of lung cancer using LIME, through three different examples. The first original lung CT image is shown, followed by a LIME superpixel heatmap, and finally, an overlay of the additives, showing the point of highest additiveness. In the first example, most of the hot superpixels are clustered in a narrow area near the central/hilar region of the lungs, and the spot is located very close to this cluster; this suggests that the model supports the malignant diagnosis more through a local focus. In the second example, the contributions are distinctly spread across both hemithoraxes. The widespread distribution of hot patches throughout the parenchyma on the LIME map and the concentration of these patches in intrapulmonary structures in the overlay image indicate that the model's decision is based on widespread tissue patterns/heterogeneity rather than a single lesion. In the third example, hot superpixels are seen to be more dominant on one side and sparser on the other. The fact that the point falls within the dominant contribution region indicates that the model gives more weight to parenchymal changes showing asymmetry for the malignant class in this section. Overall, in all three examples, the strongest contributing areas remained within the lung parenchyma, and the marked point overlapped with these areas, suggesting that the LIME description

indicates that the model's malignant classification is based not on randomness, but on the combined effect of specific parenchymal regions within the image.

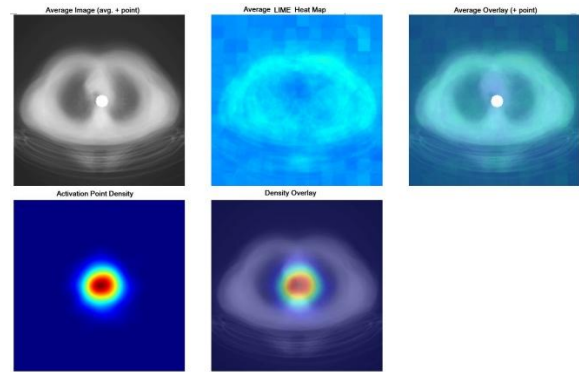


Figure 13. Average LIME activation and intensity analyses belonging to the normal class.

Figure 13 shows the LIME description generated from the average of the samples belonging to the normal class. Since no obvious findings indicating pathology are expected in the original average image, the foci seen on the LIME map do not represent the location of the disease. The model highlights the anatomical cues it references here to define the normal class. The clustering of the heatmap and density overlay panel, particularly in a region close to the center of the lungs, suggests that the model frequently checks structural features such as airway/mediastinum periphery, hilus line, and overall symmetry contrast scheme when making decisions. The clustering of activation point densities in a single core region indicates that the model repeatedly examines a similar control region across different normal samples, supporting the normal class decision through a consistent spatial pattern. Consequently, these focal points should be interpreted not as pathological lesions, but as stable anatomical reference regions that the model uses to confirm its normality.

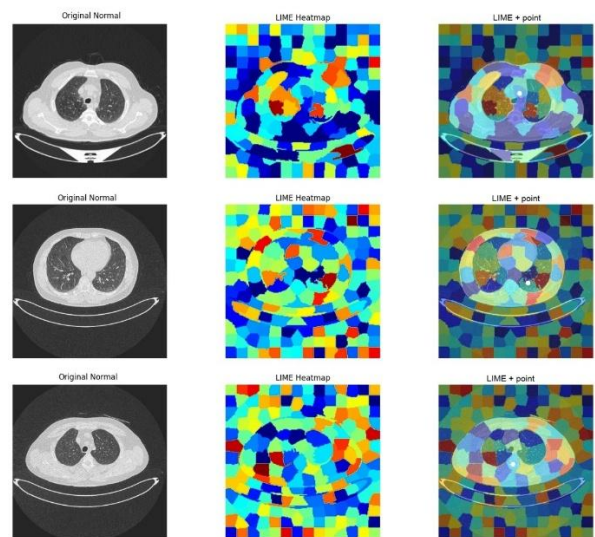


Figure 14. Example-based LIME visualizations belonging to the normal class.

Figure 14 illustrates the rationale behind the model's decision using the LIME method in three different axial thoracic CT sections belonging to the normal class. Each row contains the original image, a superpixel significance map, and an overlay of this map onto the image, along with a notation marking the point of highest contribution. Since no significant lesions are expected in the normal class, the highlighted areas reflect anatomical cues that the model confirms when making the normal decision, rather than a pathological focus. In the first example, the concentration of contributions around the upper central parenchyma and airway/bronchovascular structure suggests that the model references symmetry and typical tissue arrangement. In the second example, the peripheral contributions along the pleural line and parenchymal boundaries demonstrate that contour continuity and density transitions are distinctive for the normal class. Although the distribution is more diffuse in the third example, the fact that the highest contribution point remains in the intrapulmonary region supports the idea that the decision is based on anatomically significant areas. Overall, the model focuses not on a single focal point within the normal class, but on holistic normal anatomical features such as symmetry, borderline regularity, and typical parenchymal tissue.

5. Conclusion

This study aimed to automatically differentiate normal, benign, and malignant classes from lung CT scans. To improve the explanatory power of predictions obtained using a deep learning-based classification approach on a dataset containing a total of 1933 images, Grad-CAM++ and LIME methods were used together. The goal is not only to achieve highly accurate classification, but also to strengthen confidence in AI-based diagnostic support systems by making the model's decisions clinically interpretable.

The dataset used in this study was obtained from an existing medical image source. Therefore, the class distribution was left as is, without any changes, preserving the original structure of the dataset. The dataset contains 416 normal, 467 benign, and 1050 malignant CT images, resulting in a certain class imbalance. While such imbalances can affect model learning, different performance metrics such as precision, recall, and F1-score were used in addition to the accuracy metric to more reliably evaluate model performance.

When Grad-CAM++ outputs were evaluated on a class basis, it was observed that activations related to the normal class exhibited a low-intensity and diffuse distribution throughout the lung parenchyma, whereas the model did not tend towards a distinct pathological region. This suggests that, in normal cases, the decision is based not on the presence of a lesion, but on overall anatomical consistency supporting the absence of pathological

findings. In contrast, the concentration of activations in the malignant tumor class as more localized, asymmetrical, and distinct hotspots indicated that the model focused more selectively on parenchymal changes that might be associated with malignancy. In the benign tumor class, a more widespread and balanced distribution of attention was observed compared to the malignant class, suggesting that the model considered broader structural cues together, such as the overall tissue arrangement of the parenchyma, rather than a single aggressive focus. This separation reveals that the three classes are represented by different spatial characteristics at different levels of explainability.

Image analysis using the LIME method provided local annotations that complemented the Grad-CAM++ results. In the normal (disease-free) category, it was observed that the regions highlighted by LIME did not represent a pathological focus; the model checked normal anatomical cues such as symmetry, contour continuity, pleural border regularity, and expected density transitions when making its normal decision. In benign lung imaging samples, it was observed that the contributions remained fragmented but distributed within the parenchyma at the superpixel level, and the resolution was not reduced to a single region. In malignant lung imaging samples, it was observed that the contributions were more focused in some sections and more widespread in others, but the fact that the strongest contribution areas mostly remained within the lung parenchyma supported the anatomical consistency of the explanations. Thus, while LIME shows which regions the model's decision is locally fed from, Grad-CAM++ simultaneously reveals the class-separating distractibility of the decision-making process, demonstrating that using the two methods together significantly improves interpretability.

In conclusion, the study findings demonstrate that deep learning-based lung CT classification becomes clinically more reliable and traceable when supported not only by performance outputs but also by explainability maps. The obtained Grad-CAM++ and LIME visualizations revealed that the model primarily referenced anatomically significant intrapulmonary regions when making decisions, and the observation of different attention/contribution patterns in normal, benign, and malignant classes increased the transparency of the model's decision-making mechanism.

However, to improve the clinical generalizability of these findings, it is recommended that future external validation be performed with larger datasets from different centers, integration of 3D volumetric assessments (3D CT), analysis of class imbalance and sensitivity to imaging protocols, and quantitative comparison of XAI outputs with radiologist assessments. Accordingly, it was concluded that explainable AI-assisted classification approaches have significant potential in the early detection of lung nodules and in supporting clinical decision-making

processes.

Disclosure Statement

During the writing process of this study, generative artificial intelligence-powered tools were used to improve the text linguistically and enhance its compliance with academic writing standards. This usage does not include content creation or scientific decision-making processes.

Declaration of Ethical Standards

The authors declare that this study was conducted in accordance with accepted ethical standards in research and publication. All authors have contributed honestly to the work, appropriate citations have been provided, and the manuscript represents original research that has not been published or submitted elsewhere.

Credit Authorship Contribution Statement

All authors contributed to the conception and design of the study. Data analysis and model development were performed collaboratively. All authors participated in the interpretation of results, drafting, and critical revision of the manuscript, and approved the final version for submission.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding / Acknowledgements

No external funding was secured for this research.

Data Availability

In this study, a dataset obtained from the Mendeley platform was used. [32]. The relevant dataset can be accessed via the following link: <https://data.mendeley.com/datasets/zr4fddh833/1>

References

- [1] Q. Xiao, M. Tan, G. Yan, and L. Peng, "Revolutionizing lung cancer treatment: harnessing exosomes as early diagnostic biomarkers, therapeutics and nano-delivery platforms," *J Nanobiotechnology*, vol. 23, no. 1, p. 232, Mar 21 2025, doi: 10.1186/s12951-025-03306-0.
- [2] A. Tabur, "Akciğer Kanserlerinin Epidemiyolojik ve Klinik Özellikleri: Hastane Başvurularının Branş Dağılımlarına Göre İncelenmesi," *Journal of İzmir Chest Hospital*, pp. 26-35, 2025, doi: 10.14744/igh.2025.79188.
- [3] K. L. Hua, C. H. Hsu, S. C. Hidayati, W. H. Cheng, and Y. J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," *Onco Targets Ther*, vol. 8, pp. 2015-22, 2015, doi: 10.2147/OTT.S80733.
- [4] K. Yasufuku, "Early diagnosis of lung cancer," *Clin Chest Med*, vol. 31, no. 1, pp. 39-47, Table of Contents, Mar 2010, doi: 10.1016/j.ccm.2009.08.004.
- [5] J. Ning et al., "Early diagnosis of lung cancer: which is the optimal choice?," *Aging (Albany NY)*, vol. 13, no. 4, p. 6214, 2021. [Online]. Available: <https://www.aging-us.com/article/202504/pdf>.
- [6] N. Gómez León, S. Escalona, B. Bandrés, C. Belda, D. Callejo, and J. A. Blasco, "18F-Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography Accuracy in the Staging of Non-Small Cell Lung Cancer: Review and Cost-Effectiveness," *Radiology Research and Practice*, vol. 2014, no. 1, p. 135934, 2014, doi: 10.1155/2014/135934.
- [7] Y. Doğanşah and M. Köklü, "Use Of The Electromagnetic Spectrum In The Medical Field," *Gece Publishing*, 2022, ch. 9.
- [8] T. Kobayashi et al., "HRCT findings of small cell lung cancer measuring 30 mm or less located in the peripheral lung," *Jpn J Radiol*, vol. 33, no. 2, pp. 67-75, Feb 2015, doi: 10.1007/s11604-014-0381-2.
- [9] J. Chen et al., "Lung cancer diagnosis using deep attention-based multiple instance learning and radiomics," *Med Phys*, vol. 49, no. 5, pp. 3134-3143, May 2022, doi: 10.1002/mp.15539.
- [10] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Comput Med Imaging Graph*, vol. 31, no. 4-5, pp. 198-211, Jun-Jul 2007, doi: 10.1016/j.compmedimag.2007.02.002.
- [11] B. Van Ginneken, C. M. Schaefer-Prokop, and M. Prokop, "Computer-aided diagnosis: how to move from the laboratory to the clinic," *Radiology*, vol. 261, no. 3, pp. 719-732, 2011.
- [12] X. Zhang et al., "Deep learning enhances precision diagnosis and treatment of non-small cell lung cancer: future prospects," *Transl Lung Cancer Res*, vol. 14, no. 8, pp. 3196-3215, Aug 31 2025, doi: 10.21037/tlcr-2025-187.
- [13] L. Tian et al., "Precise and automated lung cancer cell classification using deep neural network with multiscale features and model distillation," *Sci Rep*, vol. 14, no. 1, p. 10471, May 7 2024, doi: 10.1038/s41598-024-61101-7.
- [14] Y. Zhang, Y. Weng, and J. Lund, "Applications of Explainable Artificial Intelligence in Diagnosis and Surgery," *Diagnostics (Basel)*, vol. 12, no. 2, p. 237, Jan 19 2022, doi: 10.3390/diagnostics12020237.
- [15] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, vol. 55, no. 5, pp. 3503-3568, 2022, doi: 10.1007/s10462-021-10088-y.
- [16] Z. Sadeghi et al., "A brief review of explainable artificial intelligence in healthcare," *arXiv preprint arXiv:2304.01543*, 2023, doi: 10.48550/arXiv.2304.01543.
- [17] N. Adil, P. Singh, and N. K. Nagwani, "Interpretable Lightweight CNN for Colon and Lung Cancer Classification with LIME Based Explainability," in *2024 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSTG)*, 2024: IEEE, pp. 122-127, doi: 10.1109/ICISSTG58904.2024.00034.
- [18] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097-2106, doi: 10.1109/CVPR.2017.369.
- [19] X. Li, Y. Kao, W. Shen, X. Li, and G. Xie, Lung nodule malignancy prediction using multi-task convolutional neural network (SPIE Medical Imaging). SPIE, 2017, doi: 10.1117/12.2253836.
- [20] G. Mohandass, G. Hari Krishnan, D. Selvaraj, and C. Sridhathan, "Lung Cancer Classification using Optimized Attention-based Convolutional Neural Network with DenseNet-201 Transfer Learning Model on CT image," *Biomedical Signal Processing and Control*, vol. 95, p. 106330, Sep 2024, doi: 10.1016/j.bspc.2024.106330.
- [21] M. Sivakumar, S. Chinnaamy, and T. Ms, "An efficient combined intelligent system for segmentation and classification of lung cancer computed tomography images," *PeerJ Comput Sci*, vol. 10, p. e1802, 2024, doi: 10.7717/peerj-cs.1802.
- [22] M. Budati and R. Karumuri, "An intelligent lung nodule segmentation framework for early detection of lung cancer using an optimized deep neural system," *Multimedia Tools and Applications*, vol. 83, no. 12, pp. 34153-34174, 2024, doi: 10.1007/s11042-023-17791-8.
- [23] Y. Chen, Y. Wang, F. Hu, and D. Wang, "A Lung Dense Deep Convolution Neural Network for Robust Lung Parenchyma Segmentation," *IEEE Access*, vol. 8, pp. 93527-93547, 2020,

- doi: 10.1109/access.2020.2993953.
- [24] M. Masud et al., "Light Deep Model for Pulmonary Nodule Detection from CT Scan Images for Mobile Devices," *Wireless Communications and Mobile Computing*, vol. 2020, no. 1, pp. 1-8, Jul 3 2020, doi: 10.1155/2020/8893494.
 - [25] G. Cai et al., "Medical artificial intelligence for early detection of lung cancer: A survey," *Engineering Applications of Artificial Intelligence*, vol. 159, p. 111577, Nov 8 2025, doi: 10.1016/j.engappai.2025.111577.
 - [26] P. S. H. Jose, J. E. Sagar, S. Abisheik, R. Nelson, M. Venkatesh, and B. Keerthana, "Leveraging Convolutional Neural Networks and Multimodal Imaging Data for Accurate and Early Lung Cancer Screening," in *2024 International Conference on Inventive Computation Technologies (ICICT)*, 2024: IEEE, pp. 1258-1264, doi: 10.1109/ICICT60155.2024.10544863.
 - [27] S. Singh, "Computer-aided diagnosis of thoracic diseases in chest X-rays using hybrid cnn-transformer architecture," *arXiv preprint arXiv:2404.11843*, 2024, doi: 10.48550/arXiv.2404.11843.
 - [28] N. Veeramani, A. R. S, S. P. S, S. S, and P. Jayaraman, "NextGen lung disease diagnosis with explainable artificial intelligence," *Sci Rep*, vol. 15, no. 1, p. 33052, Sep 26 2025, doi: 10.1038/s41598-025-07603-4.
 - [29] P. N. Megat Ramli, A. N. Aizuddin, N. Ahmad, Z. Abdul Hamid, and K. I. Ismail, "A Systematic Review: The Role of Artificial Intelligence in Lung Cancer Screening in Detecting Lung Nodules on Chest X-Rays," *Diagnostics (Basel)*, vol. 15, no. 3, Jan 22 2025, doi: 10.3390/diagnostics15030246.
 - [30] Z. Naz et al., "An Explainable AI-Enabled Framework for Interpreting Pulmonary Diseases from Chest Radiographs," *Cancers (Basel)*, vol. 15, no. 1, Jan 3 2023, doi: 10.3390/cancers15010314.
 - [31] A. Alghatni, T. Babu, T. R. Mahesh, S. B. Khan, O. Saidani, and M. T. Quasim, "Automated classification and explainable AI analysis of lung cancer stages using EfficientNet and gradient-weighted class activation mapping," *Front Med (Lausanne)*, vol. 12, p. 1625183, 2025, doi: 10.3389/fmed.2025.1625183.
 - [32] T. H. Talha et al., "An Image Dataset of Advanced Imaging Techniques for Lung Cancer Diagnosis," *Mendeley Data*, 2025, doi: 10.17632/xxxxx.1.
 - [33] Y. Unal, E. T. Yasin, T. A. Cengel, and M. Koklu, "Classification of Turkish hazelnut (*Corylus colurna* L.) varieties: a comparative study of YOLOv8 and fine-tuned vision transformer," *Journal of Food Measurement and Characterization*, pp. 1-15, 2026, doi: 10.1007/s11694-025-03936-w.
 - [34] E. T. Yasin et al., "Detecting Driver Fatigue Using Artificial Intelligence on a Realistic Driving Images," *Journal of Future Artificial Intelligence and Technologies*, vol. 2, no. 4, pp. 648-660, 2026, doi: 10.62411/faith.3048-3719-299.
 - [35] I. A. Ozkan and M. Koklu, "Skin lesion classification using machine learning algorithms," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 5, no. 4, pp. 285-289, 2017, doi: 10.18201/ijisae.2017534420.
 - [36] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging," *Cancers (Basel)*, vol. 15, no. 16, p. 4172, Aug 18 2023, doi: 10.3390/cancers15164172.
 - [37] M. M. Saritas, O. Kilci, and M. Koklu, "Evaluation of CNN Models for Multi-Class Gear Fault Detection Using Waveform Images," in *Proceedings of International Conference on Intelligent Systems and New Applications*, 2025, vol. 3, pp. 31-40, doi: 10.58190/icisna.2025.137.
 - [38] I. Cinar and M. Koklu, "Identification of rice varieties using machine learning algorithms," *Journal of Agricultural Sciences*, pp. 9-9, 2022, doi: 10.15832/ankutbd.862482.
 - [39] F. Haque et al., "An End-to-End Concatenated CNN Attention Model for the Classification of Lung Cancer With XAI Techniques," *IEEE Access*, vol. 13, pp. 96317-96336, 2025, doi: 10.1109/access.2025.3572423.
 - [40] M. Koklu, R. Kursun, Y. S. Taspinar, and I. Cinar, "Classification of date fruits into genetic varieties using image analysis," *Mathematical Problems in Engineering*, vol. 2021, no. 1, p. 4793293, 2021, doi: 10.1155/2021/4793293.
 - [41] H. H. Aras, Y. Eryeşil, and M. Köklü, "An Explainable Deep Learning Framework for Agron-Based Coffee Roast Classification Using Grad-CAM," in *Proceedings of International Conference on Intelligent Systems and New Applications*, 2025, vol. 3, pp. 51-57, doi: 10.58190/icisna.2025.139.
 - [42] B. Işgor and M. Koklu, "Lightweight Hybrid Model for Bone Fracture Detection Using MobileNetV2 Feature Extraction and Ensemble Learning," *Journal of Future Artificial Intelligence and Technologies*, vol. 2, no. 3, pp. 521-533, 2025, doi: 10.62411/faith.3048-3719-284.
 - [43] H. Incekara, I. H. Cizmeci, M. M. Saritas, and M. Koklu, "Classification of almond kernels with optuna hyper-parameter optimization using machine learning methods," *Journal of Food Science and Technology*, pp. 1-17, Nov 25 2025, doi: 10.1007/s13197-025-06494-7.
 - [44] İ. H. Çizmeci and H. Incekara, "Stacking Ensemble Based Hybrid Machine Learning Approach for Predicting Obesity Levels," in *2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*, 2025: IEEE, pp. 1-8, doi: 10.1109/ICHORA65333.2025.11017164.
 - [45] D. Garreau and U. Luxburg, "Explaining the explainer: A first theoretical analysis of LIME," in *International Conference on Artificial Intelligence and Statistics*, 2020: PMLR, pp. 1287-1296.
 - [46] M. D. İli and F. Özyurt, "Açıklanabilir Yapay Zeka Yöntemleriyle Mr Görüntülerinden Beyin Tümörü Tespiti," *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 28, no. 2, pp. 1092-1109, 2025.
 - [47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.
 - [48] D. Theckedath and R. Sedamkar, "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks," *SN Computer Science*, vol. 1, no. 2, p. 79, 2020, doi: 10.1007/s42979-020-0114-9.



Research Article

Domain-Specific Fine-Tuning of LLMs for Elevator Emergency Response Systems

Busra Onal^a , Muhammet Fatih Aslan^{b,*} , Akif Durdu^c ^a Butkon Elevator Industry Trade Inc., Research and Development Center, Konya, Türkiye^b Department of Artificial Intelligence and Machine Learning, Faculty of Computer and Information Sciences, Konya Technical University, Konya, Türkiye^c Department of Electrical-Electronic Engineering, Konya Technical University, Konya, Türkiye

ARTICLE INFO

Article history:

Received 3 February 2026

Accepted 25 March 2026

Keywords:

Edge AI,
Elevator Safety,
Fine-Tuning,
Large Language Models,
QLoRA.

ABSTRACT

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation. However, their application in safety-critical domains such as elevator emergency response requires domain-specific knowledge and reliable performance, particularly when deployed as offline edge AI systems with constrained computational resources. This study presents fine-tuning of the Gemma-2-9B-Instruct model using Quantized Low-Rank Adaptation (QLoRA) for Turkish elevator emergency scenarios, targeting deployment on embedded touchscreen control panels inside elevator cabins. Unlike cloud-based systems leveraging internet connectivity for retrieval, our edge deployment operates entirely offline, making fine-tuning essential for encoding domain knowledge directly into model parameters. We developed a specialized Turkish dataset containing 1,155 question-answer pairs covering diverse emergency situations. Our evaluation demonstrates significant improvements: ROUGE-1 scores increased from 0.259 to 0.317 (22.49%), BLEU improved by 218.11%, and hallucination rates reduced from 54% to 22% while achieving 49% faster inference. The resulting 4.8GB quantized model runs entirely on embedded hardware without network dependencies, providing immediate, reliable emergency guidance. These results validate parameter-efficient fine-tuning for safety-critical edge AI applications.

This is an open access article under the CC BY-SA 4.0 license.
(<https://creativecommons.org/licenses/by-sa/4.0/>)

1. Introduction

Large Language Models (LLMs) have significantly advanced natural language processing by enabling machines to generate coherent, context-aware, and semantically meaningful text across a wide range of tasks. These capabilities are primarily driven by transformer-based architectures and large-scale pretraining on heterogeneous textual corpora, which allow models to capture linguistic structure, semantic relationships, and contextual dependencies across domains [1, 2]. Subsequent developments in instruction tuning and alignment techniques have further improved the ability of LLMs to follow user intent and produce structured

responses suitable for interactive applications [3]. With the emergence of open-weight model families such as LLaMA [4] architectures and the Gemma [5] series, the deployment of LLM-based intelligent assistants has expanded beyond research prototypes into practical industrial experimentation. Nevertheless, most existing research and deployment scenarios remain focused on general-purpose conversational systems rather than domain-constrained environments where reliability, response structure, and deterministic behavior are essential.

Deploying LLMs in safety critical industrial environments introduces challenges that differ fundamentally from those encountered in conventional

* Corresponding author. E-mail address: mfaslan@ktun.edu.tr
DOI: 10.58190/ijamec.2026.165

conversational AI systems. In such environments, incorrect or misleading responses may create operational risks rather than minor usability problems. One of the most widely documented reliability issues in generative language models is hallucination, where models produce fluent but factually incorrect information with high confidence [6]. This phenomenon becomes particularly problematic when models are applied to specialized technical domains or low-resource language settings, where domain terminology, procedural knowledge, and regulatory information are insufficiently represented in pretraining corpora[7]. Industrial safety applications require communication that is not only linguistically correct but also procedurally accurate, concise, and actionable. Consequently, ensuring factual reliability, predictable response structure, and domain-appropriate communication becomes a central requirement when integrating LLM based assistants into safety-critical systems.

Elevator emergency response systems provide a representative example of such constrained interaction scenarios. Conventional emergency communication mechanisms (including alarm buttons, printed instructions, and intercom systems) offer static guidance and cannot dynamically adapt to user uncertainty during stressful situations. During emergency events such as entrapment, fire, earthquake, or medical incidents, occupants may experience panic, reduced situational awareness, and difficulty interpreting written instructions. An embedded AI assistant capable of delivering structured emergency guidance in natural language could improve safety outcomes by providing immediate procedural instructions and reassurance. However, implementing such functionality requires addressing deployment constraints that differ substantially from those assumed in typical cloud-based AI applications.

Unlike most contemporary LLM deployments, elevator cabin systems must operate as offline edge AI solutions. Elevators frequently function in connectivity-limited environments such as building shafts, underground parking areas, and reinforced structural cores, where wireless communication may be unreliable or unavailable. Emergency conditions may further involve communication failures, making reliance on external services unacceptable. At the same time, embedded cabin control panels operate under strict hardware limitations, typically providing only a few gigabytes of memory and low-power processors. These constraints make conventional cloud inference pipelines and retrieval-dependent architectures impractical, motivating research on parameter-efficient adaptation, quantization, and compact deployment strategies that enable fully self-contained AI assistants.

Retrieval-Augmented Generation (RAG) has emerged as a widely adopted approach for improving factual

consistency in LLM-based systems by grounding responses in external knowledge sources [8]. By retrieving relevant documents during inference, RAG systems can reduce hallucination and improve factual alignment with authoritative information. However, retrieval pipelines introduce additional latency, memory overhead, and infrastructure dependencies that may conflict with real-time safety-critical interaction requirements on embedded hardware. Furthermore, fragmented user queries typical of panic situations may reduce retrieval effectiveness. These limitations motivate investigation of alternative approaches that maintain reliability while enabling deterministic and fully offline operation in resource-constrained environments.

Fine-tuning is a widely used approach for adapting pretrained large language models to downstream tasks by updating model parameters using task-specific data. Rather than training models from scratch, fine-tuning allows previously learned linguistic and semantic representations to be reused while specializing model behavior for new applications. Parameter-efficient fine-tuning methods, such as Low-Rank Adaptation (LoRA) [9], enable this adaptation with minimal computational cost by updating only a small subset of model parameters. Quantized Low-Rank Adaptation (QLoRA) [10] further extends this idea by combining LoRA with low-precision quantization, allowing large models to be fine-tuned and deployed under strict memory constraints while maintaining performance. A specialized form of this process, known as domain-specific fine-tuning, focuses on adapting language models to a particular knowledge domain by embedding domain terminology, procedural knowledge, and communication patterns directly into model parameters. Prior research in specialized domains (BioBERT [11], LEGAL-BERT [12], etc.) demonstrates that domain-adapted models can achieve improved terminology usage, factual reliability, and task performance compared to general-purpose LLMs. These findings suggest that parameter-efficient domain-specific fine-tuning is particularly well suited for safety-critical edge AI applications, where computational efficiency and domain accuracy must be balanced.

In this work, we investigate QLoRA-based fine-tuning of the Gemma-2-9B-Instruct model for Turkish elevator emergency response scenarios, targeting deployment on embedded touchscreen control panels inside elevator cabins. A domain-specific dataset consisting of 1,155 Turkish question-answer pairs was constructed using elevator safety manuals, incident reports, and expert validation. The adapted model is deployed using 4-bit quantization, enabling fully offline execution on resource-constrained hardware[10]. Experimental results demonstrate improvements in response quality, reduction in hallucination rates, and faster inference compared to the base model, indicating that parameter-efficient domain

adaptation can enable reliable LLM deployment in safety-critical industrial environments. The primary contributions of this work are:

Development of a domain-specific Turkish elevator emergency response dialogue dataset covering diverse real-world emergency scenarios, validated by certified safety professionals against international standards.

Parameter-efficient fine-tuning of a 9B-parameter instruction-tuned LLM using QLoRA, enabling effective domain adaptation while updating less than 1% of model parameters.

Offline deployment of a fully quantized LLM on embedded hardware, eliminating network dependency for safety-critical real-time emergency guidance.

A comprehensive evaluation combining generation-quality, reliability, and efficiency metrics, including hallucination rate assessment tailored for safety-critical applications.

2. Related Work

LLMs often require adaptation to specialized domains in order to achieve reliable performance. Although general-purpose models are trained on large and diverse corpora, they may struggle with domain-specific terminology due to domain shift in specialized texts. To address this problem, prior research has adopted domain-specific pre-training and fine-tuning strategies. For example, BioBERT[11], developed for biomedical text mining, was trained on biomedical corpora and demonstrated improved performance over the general BERT model in named entity recognition (NER) and relation extraction (RE) tasks. Similarly, SciBERT[13] achieved performance improvements for scientific text processing by introducing a domain-specific vocabulary (SCIVOCAB). ClinicalBERT[14], trained on clinical notes, and LEGAL-BERT[12], designed for legal documents, further demonstrate the importance of domain adaptation in specialized language modeling. In the financial domain, FinBERT[15] has shown improved performance in financial sentiment analysis compared to general-purpose language models.

To adapt large language models efficiently, Parameter-Efficient Fine-Tuning (PEFT) methods have gained popularity as an alternative to full-model fine-tuning. LoRA reduces training cost by freezing pretrained model weights and injecting trainable low-rank matrices into transformer layers, significantly reducing the number of trainable parameters. A more recent approach, QLoRA, further improves memory efficiency by combining LoRA with low-precision quantization. QLoRA introduces innovations such as the 4-bit NormalFloat data type and double quantization, enabling fine-tuning of models with tens of billions of parameters on limited hardware without significant performance degradation.

Research on the use of LLMs in elevator systems remains limited, although recent studies have begun to explore maintenance and technical-support applications. A recent study focusing on elevator maintenance strategies demonstrated that large language models can be adapted using domain-specific maintenance data and that LoRA-based fine-tuning can improve diagnostic accuracy [16]. The study emphasizes that reliable information generation is critical in maintenance workflows and that hallucination remains an important safety concern. Another study investigated the use of GPT-4-based LLMs as evaluation tools for task-oriented industrial chatbots [17]. The results indicate that LLM-based evaluation aligns well with human judgment for low-complexity tasks but becomes less reliable for higher cognitive complexity scenarios. These findings highlight the importance of reliability assessment when deploying LLM-based systems in industrial environments.

Existing literature indicates that LLM applications in elevator systems primarily focus on maintenance optimization and product-selection assistance. However, offline emergency-guidance assistants operating inside elevator cabins on embedded hardware remain largely unexplored. This study aims to address this gap by developing an elevator emergency assistant capable of running on resource constrained hardware using the memory efficient QLoRA fine-tuning approach, enabling reliable emergency guidance without requiring internet connectivity.

3. Methodology

This study aims to develop a domain-adapted large language model capable of providing reliable emergency guidance within elevator cabins under offline and hardware-constrained conditions. Unlike cloud-based conversational systems, the proposed approach focuses on embedding procedural knowledge directly into model parameters through parameter-efficient fine-tuning. The methodological framework is designed to ensure three primary objectives: (i) domain specialization for Turkish elevator emergency scenarios, (ii) memory-efficient adaptation compatible with embedded deployment, and (iii) reduction of hallucination risks in safety-critical responses.

The overall methodology consists of four sequential stages. First, a structured domain-specific dataset was constructed based on realistic elevator emergency scenarios. Second, a pretrained instruction-tuned large language model (Gemma2-9B-Instruct) was selected as the base model. Third, parameter-efficient fine-tuning was applied using the QLoRA method to adapt the model to the emergency domain. Finally, the trained model was evaluated and prepared for deployment in a quantized format suitable for embedded systems. The following

subsections describe each stage in detail.

3.1. Dataset Construction

A domain-specific Turkish dataset consisting of 1,155 question-answer pairs was developed to represent realistic elevator emergency scenarios. Data collection involved a systematic review of safety manuals from major elevator manufacturers (Otis, Schindler, KONE, ThyssenKrupp), consultation with certified elevator safety professionals, and analysis of documented incident reports. The objective was to construct a dataset that reflects procedurally accurate, safety-compliant, and context-aware emergency guidance suitable for deployment in elevator cabin environments.

The dataset covers seven primary categories: (1) mechanical failures, (2) entrapment situations, (3) fire and smoke incidents, (4) earthquake-related scenarios, (5) medical emergencies, (6) safety concerns such as air supply and structural integrity, and (7) communication failures. Each response was written in a concise, directive format prioritizing clarity, actionability, and reassurance, critical characteristics for crisis communication. All samples follow a structured instruction, response format compatible with instruction-tuned transformer models. A representative example is shown in Figure 1.



Figure 1. Sample instruction-response structure used for fine-tuning (Turkish original (left) and English translation (right).)

To ensure dataset reliability and safety compliance, a multi-stage quality assurance process was applied. This included expert review by certified elevator safety professionals, linguistic quality control by native Turkish speakers, automated consistency checks through validation scripts, and safety verification against international standards, including EN 81 and ASME A17.1. These procedures were implemented to minimize ambiguity, eliminate procedural inconsistencies, and reduce the risk of unsafe or misleading guidance.

The dataset was divided into training and testing subsets using stratified sampling to preserve category balance. An 85%/15% split was applied, resulting in 1,039 training samples and 116 test samples. This stratified approach ensured that all emergency categories were proportionally represented in both subsets, supporting fair evaluation of

model performance.

3.2. Model: Gemma-2-9B-Instruct

As the base model for domain adaptation, Gemma-2-9B-Instruct [5] was selected due to its balance between performance, architectural efficiency, and compatibility with parameter-efficient fine-tuning methods. Gemma-2 is a transformer-based large language model trained using large scale corpora and subsequently instruction-tuned to improve task-following and conversational alignment. The 9-billion-parameter variant offers sufficient representational capacity for domain adaptation while remaining feasible for quantized deployment on resource-constrained hardware.

The model follows a decoder-only transformer architecture with multi-head self-attention layers and feed-forward networks optimized for autoregressive text generation. Its instruction-tuned configuration allows it to interpret structured prompt formats, including role-based conversational inputs, making it suitable for the emergency guidance dialogue structure introduced in Section 3.1. Since the dataset was constructed in a user-assistant message format, alignment with the model's original instruction-tuning objective was preserved during adaptation.

One of the key reasons for selecting Gemma-2-9B-Instruct instead of larger-scale models (e.g., 30B+ parameters) was deployment feasibility. The target system is an embedded elevator cabin control panel with strict memory and computational constraints. Larger models would require either multi-GPU infrastructure or aggressive compression strategies that could negatively impact reliability. Conversely, smaller models (e.g., <3B parameters) may lack sufficient representational capacity to internalize procedural safety instructions effectively. Therefore, the 9B configuration represents a compromise between domain adaptation capability and edge deployment practicality.

Before fine-tuning, the pretrained weights were loaded in 4-bit precision format to enable memory-efficient training using QLoRA. The base model parameters remained frozen during adaptation, ensuring that domain knowledge was incorporated exclusively through lightweight adapter layers. This design choice reduces catastrophic forgetting and preserves the general linguistic competence of the pretrained model while injecting elevator-specific emergency knowledge.

3.3. QLoRA Configuration

To enable memory-efficient domain adaptation of the Gemma-2-9B-Instruct model, QLoRA[10] was employed. The complete configuration and resource footprint are summarized in Table 1. As shown in Table 1, the pretrained model weights were loaded in 4-bit NormalFloat (NF4) precision with double quantization

enabled, while training computations were performed in bfloat16 format to ensure numerical stability. Instead of updating all model parameters, LoRA adapters were injected into the transformer attention projection layers, specifically the `q_proj`, `k_proj`, `v_proj`, and `o_proj` modules.

Table 1. QLoRA Configuration Parameters

Category	Parameter	Value
Base Model	Model Name	Gemma-2-9B-Instruct
Quantization	Weight Precision	4-bit NF4
	Double Quantization	Enabled
	Compute Data Type	bfloat16
LoRA Configuration	Rank (r)	32
	Alpha (α)	64
	Dropout	0.05
Target Modules	Injected Layers	<code>q_proj</code> , <code>k_proj</code> , <code>v_proj</code> , <code>o_proj</code>
Training Output	Trainable Parameters	37M (~0.4%)
Memory Footprint	Adapter Size	148 MB
Deployment Size	Final Quantized Model	4.8 GB

The LoRA configuration was defined with a rank (r) of 32, a scaling factor (α) of 64, and a dropout rate of 0.05. Under this setup, only 37 million parameters were updated during training, corresponding to approximately 0.4% of the total model parameters. The resulting adapter size was 148 MB, while the fully quantized deployed model occupied 4.8 GB of memory. This configuration significantly reduces memory requirements compared to full fine-tuning while preserving sufficient adaptation capacity for domain-specific learning. As indicated in Table 1, the final model size and adapter footprint make the system suitable for deployment in resource-constrained embedded environments such as elevator cabin control units.

3.4. Training

The model was fine-tuned using the domain-specific dataset described in Section 3.1 and the QLoRA configuration presented in Table 1. The complete training setup and outcomes are summarized in Table 2. During preprocessing, the Gemma-2 chat template was applied to maintain alignment with the instruction-tuned architecture of the base model. Label masking was enabled to ensure that loss was computed only over assistant response tokens, preventing unintended optimization over user prompts. Sequences were truncated to a maximum length of 2,048 tokens, with 98.7% of samples naturally fitting within this limit, indicating minimal truncation impact on

training quality.

Training was performed using Paged AdamW (32-bit) optimizer with a learning rate of 1×10^{-4} . A cosine learning rate schedule was applied, including 50 warmup steps to stabilize early training dynamics. Weight decay was set to 0.01, and gradient clipping at 0.3 was applied to prevent instability during backpropagation. Mixed precision training was conducted using bfloat16 computation to balance numerical stability and memory efficiency.

Due to memory constraints and to reflect deployment-oriented conditions, a batch size of 2 was used. Gradient accumulation of 16 steps resulted in an effective batch size of 32. Training was conducted for 5 epochs, totaling 165 optimization steps. Model evaluation was performed every 50 steps, and early stopping with a patience of 3 evaluations was applied to prevent overfitting. Training was executed on a single NVIDIA A100 GPU, requiring approximately 43 minutes to complete.

Table 2. Training Configuration and Outcomes

Category	Parameter	Value
Preprocessing	Chat Template	Gemma-2 template
	Label Masking	Enabled
	Max Sequence Length	2048 tokens
	Natural Fit Rate	98.7%
Optimization	Optimizer	Paged AdamW (32-bit)
	Learning Rate	1×10^{-4}
	Scheduler	Cosine
	Warmup Steps	50
	Gradient Clipping	0.3
	Weight Decay	0.01
Training Setup	Batch Size	2
	Gradient Accumulation	16
	Effective Batch Size	32
	Epochs	5
	Total Steps	165
	Mixed Precision	bfloat16
	Evaluation Interval	Every 50 steps
	Early Stopping Patience	3
Hardware	GPU	NVIDIA A100
	Training Time	43 minutes
Outcome	Final Validation Loss	1.4775
	Perplexity	4.38
	Improvement vs Base	44.7%

As reported in Table 2, the final validation loss achieved was 1.4775, corresponding to a perplexity of 4.38. Compared to the base model configuration, this represents a 44.7% improvement in validation performance, indicating successful domain adaptation under parameter-efficient constraints.

3.5. Evaluation

To assess the effectiveness of the proposed domain-

adapted model, evaluation was conducted using three complementary perspectives: automatic text generation metrics, domain-specific reliability measures, and system-level performance analysis. Automatic evaluation metrics were used to quantify lexical and semantic similarity between generated responses and reference answers. ROUGE[18] measures token-level overlap, including unigram (ROUGE-1), bigram (ROUGE-2), and longest common subsequence (ROUGE-L) scores. BLEU[19] evaluates n-gram precision with a brevity penalty mechanism. METEOR[20] incorporates synonym matching, stemming, and word order considerations. BERTScore[21] computes contextual semantic similarity using pretrained multilingual BERT embeddings. These metrics provide quantitative indicators of response alignment with expert-written references.

Given the safety-critical nature of elevator emergency guidance, automatic metrics alone are insufficient. Therefore, domain-specific evaluation criteria were introduced. Faithfulness measures factual consistency with established safety procedures. A subset of 100 randomly selected test samples was annotated by certified professionals to assess technical accuracy. Answer Correctness evaluates whether responses directly address the user query and provide actionable guidance. Hallucination Rate identifies uncertain expressions (e.g., speculative language), fabricated specifications, or contradictory instructions that may compromise safety.

System-level performance was also evaluated. Latency was measured as average generation time per response, while throughput was calculated as the number of queries processed per second. Token statistics were analyzed to assess generation efficiency. All evaluations were conducted on an NVIDIA A100 GPU using identical generation parameters: temperature 0.1, top-p 0.95, and maximum output length of 256 tokens.

4. Results

4.1. Training and Quality Evaluation

Fine-tuning converged smoothly over five epochs, achieving a final validation loss of 1.4775 (perplexity 4.38), compared to the base model's perplexity of 7.92, indicating substantial domain adaptation. The comprehensive evaluation comparing the base and fine-tuned models is presented in Table 3.

Table 3. Comprehensive Evaluation: Base vs. Fine-Tuned Model

Metric	Base	Fine-Tuned	Abs. Δ	Rel. Δ (%)
ROUGE-1	0.2586	0.3168	+0.0582	+22.49
ROUGE-2	0.0716	0.1106	+0.0389	+54.36
ROUGE-L	0.1525	0.1986	+0.0461	+30.19
BLEU	0.0131	0.0417	+0.0286	+218.11
METEOR	0.1542	0.1759	+0.0217	+14.06
BERTScore F1	0.6559	0.7050	+0.0491	+7.49
Correctness (%)	54.49	60.49	+6.00	+11.01
Hallucination (%)	54.00	22.00	-32.00	-59.26
Avg Latency (s)	19.72	10.05	-9.67	-49.04
Throughput (s/s)	0.051	0.099	+0.049	+96.24
Avg Output Tokens	238.95	121.51	-117.44	-49.15

As shown in Table 3, the fine-tuned model consistently outperformed the base model across all automatic evaluation metrics. ROUGE-1 improved by 22.49%, indicating stronger lexical alignment with domain-specific terminology. ROUGE-2 showed a significant improvement of 54.36%, reflecting enhanced bigram precision and better phrase-level structuring in emergency instructions. ROUGE-L increased by 30.19%, suggesting improved sequence-level coherence.

BLEU exhibited the largest relative improvement (+218.11%), highlighting substantial gains in n-gram precision. Although BLEU scores remain numerically lower than ROUGE values, as typical in instruction-based generation tasks, the relative gain demonstrates stronger alignment with reference safety responses. METEOR improved by 14.06%, and BERTScore F1 increased by 7.49%, indicating enhanced semantic similarity and contextual alignment.

Beyond automatic metrics, domain-specific evaluation revealed critical improvements in procedural reliability. Answer correctness increased from 54.49% to 60.49%, while hallucination rate decreased dramatically from 54% to 22%, representing a 59.26% reduction. This reduction is particularly significant given the safety-critical context of elevator emergency guidance. The base model frequently produced uncertain or speculative language, whereas the fine-tuned model generated more grounded and directive responses aligned with established safety procedures.

Fine-tuning improved not only response quality but also system efficiency. Average latency decreased from 19.72 seconds to 10.05 seconds (-49.04%), largely due to shorter and more focused outputs. The average output length was reduced from 238.95 tokens to 121.51 tokens, reflecting improved response conciseness. Throughput nearly doubled, increasing by 96.24%, demonstrating that domain adaptation enhanced both computational efficiency and response clarity.

These results indicate that parameter-efficient domain-specific fine-tuning successfully improved lexical

alignment, semantic consistency, procedural correctness, and runtime performance simultaneously, an important requirement for embedded safety-critical applications. The radar visualization in Figure 2 provides a consolidated view of the performance differences between the base and fine-tuned models. The expanded area covered by the fine-tuned model indicates consistent improvements across lexical, semantic, reliability, and efficiency metrics. This visual summary reinforces the quantitative findings reported in Table 3.

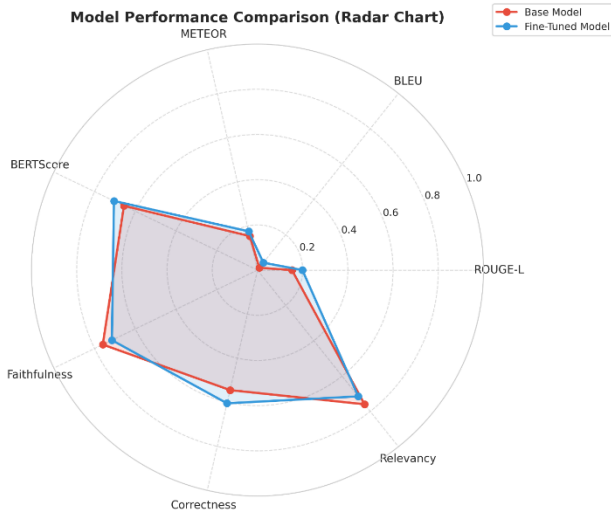


Figure 2. Radar chart comparing models. Fine-tuned model (blue) shows consistent improvements across metrics.

The radar chart presented in Figure 2 provides a multidimensional view of the performance differences between the base and fine-tuned models across six key evaluation dimensions: BLEU, ROUGE-L, Correctness, Relevancy, Faithfulness, and BERTScore. The fine-tuned model consistently occupies a larger area across all axes, confirming that improvements are not confined to a single dimension but are distributed uniformly across lexical, semantic, and reliability-oriented criteria. Notably, the most pronounced expansions are observed in the Faithfulness and Correctness axes, reflecting the model's improved alignment with established safety procedures and its ability to provide procedurally accurate responses. The BERTScore and ROUGE-L axes also demonstrate meaningful gains, indicating stronger contextual coherence and sequence-level alignment with reference responses. In contrast, the base model exhibits a considerably smaller and uneven coverage across the radar dimensions, suggesting inconsistent performance that is particularly inadequate for safety-critical deployment scenarios. Taken together, the radar visualization reinforces the quantitative findings reported in Table 3 and highlights that domain-specific fine-tuning produces a balanced and reliable model suitable for embedded emergency guidance applications.

4.2. Qualitative Example

To further illustrate the behavioral differences between the base and fine-tuned models, a representative emergency query was evaluated.

Query (Turkish): “Asansörde sıkıştım, kapı açılmıyor, ne yapmalıyım?” (“I am stuck in the elevator, the door will not open, what should I do?”)

Base Model Response (Turkish): “Öncelikle sakin olmaya çalışın. Asansörlerde genellikle acil durum butonları bulunur. Yardım istemek için bu butonları kullanabilirsiniz. Kapıyı zorlamamanız önerilir.” (“First, try to remain calm. Elevators generally have emergency buttons. You can use these buttons to request help. It is recommended not to force the door.”) The base model response is relatively generic and includes hedging expressions such as “genellikle” (“generally”). The instructions lack procedural sequencing and do not provide specific details such as button identification or expected rescue timeframes. The tone is cautious but not directive.

Fine-Tuned Model Response (Turkish): “Sakin kalın ve panik yapmayın. Alarm butonuna (kırmızı veya zil işaretli) basarak yardım çağırın. Kapıyı zorlamayın, tehlikelidir. Kabin havalandırması aktiftir. 112’yi arayın. Yardım 15–30 dakika içinde gelir. Oturup bekleyin.” (“Remain calm and do not panic. Press the alarm button (red or marked with a bell symbol) to call for assistance. Do not force the door, it is dangerous. Cabin ventilation is active. Call emergency services (112). Help will arrive within 15–30 minutes. Sit down and wait.”) The fine-tuned model produces a structured and sequential response with explicit procedural guidance. The instructions are directive, include identifiable physical cues (button appearance), and provide a time expectation for rescue. The language avoids uncertainty and maintains a confident tone, which is essential in crisis communication scenarios.

The qualitative comparison demonstrates that domain-specific fine-tuning substantially improves structural clarity, procedural specificity, and confidence calibration. While the base model produces a generally relevant response, it relies on generic phrasing and hedging expressions, resulting in less decisive guidance. In contrast, the fine-tuned model delivers instructions in a clear sequential order, ensuring that users can easily follow step-by-step actions during a stressful situation. The inclusion of actionable details, such as identifying the alarm button by its color or symbol and providing an estimated rescue timeframe, reflects the internalization of procedural knowledge during fine-tuning. Furthermore, the elimination of speculative language enhances decisiveness, which is particularly critical in emergency communication. These differences indicate that parameter-efficient domain adaptation not only improves lexical alignment and metric-based performance but also meaningfully enhances response reliability and practical usability in safety-critical elevator scenarios.

5. Discussion

The findings of this study demonstrate that parameter-efficient fine-tuning via QLoRA enables effective adaptation of large language models to safety-critical elevator emergency scenarios under embedded hardware constraints. As illustrated in Figure 3, the fine-tuned model consistently outperforms the base model across all automatic evaluation metrics, including ROUGE, BLEU, METEOR, and BERTScore. The visual comparison confirms that performance gains are not isolated to a single metric but reflect balanced improvements in lexical alignment, phrase-level structuring, and semantic similarity. This pattern indicates that domain-specific fine-tuning modifies not only vocabulary usage but also response structure and contextual coherence.

Also, findings are consistent with prior domain adaptation studies in other specialized fields. Similar to BioBERT and LEGAL-BERT, which demonstrated that embedding domain-specific knowledge into model parameters substantially improves task performance over general-purpose models, our results confirm that this principle extends to safety-critical industrial applications in low-resource language settings. Furthermore, the observed reduction in hallucination rate aligns with findings reported in elevator maintenance LLM studies, where factual reliability was identified as a primary concern in industrial deployments. The simultaneous gains in both response quality and inference efficiency also complement recent observations in parameter-efficient fine-tuning literature, suggesting that QLoRA-based adaptation represents a broadly viable strategy for resource-constrained domain specialization beyond the specific scenario investigated here.

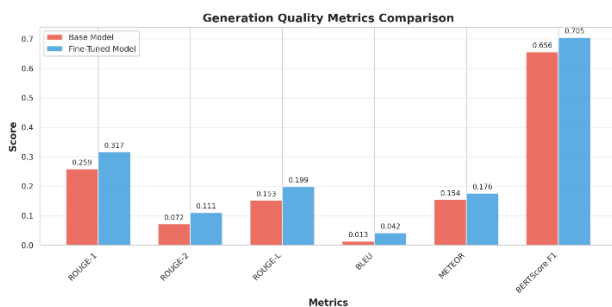


Figure 3. Generation quality metrics showing consistent gains across ROUGE, BLEU, METEOR, and BERTScore.

Beyond lexical and semantic gains, the most significant improvement concerns reliability. The substantial reduction in hallucination rate suggests that domain adaptation constrains generative variability and promotes more deterministic, grounded responses. In safety-critical environments such as elevator cabins, speculative phrasing or uncertain language can directly influence user behavior during emergencies. The fine-tuned model demonstrates clearer sequential structuring, directive phrasing, and confidence calibration, reflecting internalization of

procedural emergency knowledge. These qualitative shifts align with the quantitative improvements observed in Table 3 and visually summarized in Figure 3.

Another notable outcome is the simultaneous enhancement of efficiency. The fine-tuned model generates shorter, more focused responses, resulting in reduced latency and improved throughput. This is particularly important for edge AI deployment, where computational resources are limited and response time is critical. Unlike many adaptation strategies that improve accuracy at the cost of increased computation, the proposed configuration improves both quality and runtime performance, making it well suited for embedded emergency systems.

From a system architecture perspective, these results support the selection of fine-tuning over retrieval-augmented generation (RAG) for this deployment scenario. While RAG offers flexibility in updating knowledge dynamically, it introduces reliance on network connectivity and additional latency. Elevator cabins (especially during emergencies) may lack stable connectivity, and deterministic offline operation becomes mandatory. Embedding procedural knowledge directly into model parameters through fine-tuning therefore provides a more robust and deployment-feasible solution for this context.

Despite the observed improvements, several limitations remain. The dataset, although expert-validated, may not fully cover rare or high-severity edge cases such as complex multi-person emergencies or cascading mechanical failures. Automatic evaluation metrics, while informative, cannot fully capture the appropriateness and contextual safety of responses. Furthermore, the residual hallucination rate indicates that human oversight mechanisms remain necessary. Techniques such as confidence thresholding, structured output templates, human-in-the-loop escalation for complex scenarios, and periodic expert auditing should be integrated into real-world deployments to ensure responsible use.

Although the dataset includes a number of short and fragmented queries reflecting realistic user behavior under stress, the construction process did not specifically prioritize panic language patterns such as broken sentences, single-word exclamations, or incoherent phrasing. Emergency situations frequently elicit such atypical linguistic inputs, and the model's robustness to these patterns was not systematically evaluated in the current study. Future work should incorporate a dedicated subset of panic-language samples into the training data, alongside targeted evaluation protocols designed to assess model performance under degraded input conditions.

Overall, the results suggest that quantization-aware, parameter-efficient fine-tuning provides a viable pathway for integrating LLM-based assistants into embedded safety-critical systems. By balancing procedural

reliability, semantic alignment, and computational efficiency, the proposed approach advances the practical application of large language models in real-world emergency response environments.

6. Conclusion

This study demonstrates that parameter-efficient fine-tuning using QLoRA enables effective adaptation of large language models to safety-critical edge AI applications. By fine-tuning Gemma2-9B-Instruct on a domain-specific Turkish elevator emergency dataset, substantial improvements were achieved across multiple dimensions, including generation quality, semantic alignment, procedural correctness, hallucination reduction, and runtime efficiency. Notably, these gains were obtained while updating only 37 million parameters (0.4% of the model) and maintaining a deployment-ready 4.8 GB quantized footprint.

The results confirm that domain-specific fine-tuning does more than improve lexical overlap; it reshapes response structure, confidence calibration, and procedural grounding, key attributes for emergency communication systems. The significant reduction in hallucination rate further underscores the importance of embedding domain knowledge directly into model parameters when operating in safety-critical contexts. From a deployment perspective, the proposed approach validates the feasibility of running large language models entirely offline on embedded hardware. Unlike retrieval-based architectures, the fine-tuned model operates deterministically without external network dependency, making it suitable for environments where connectivity cannot be guaranteed.

The practical implications of this work extend beyond the specific elevator emergency domain. The proposed pipeline (combining domain-specific dataset construction, parameter-efficient QLoRA fine-tuning, and 4-bit quantized deployment) provides a replicable framework applicable to a wide range of safety-critical edge AI scenarios, including industrial machinery operation, mining safety systems, and disaster response guidance. The demonstrated ability to reduce hallucination rates by 59.26% while simultaneously improving inference speed by 49% underscores that reliability and efficiency are not mutually exclusive objectives in embedded AI design. Furthermore, the relatively modest hardware requirements and short training time of 43 minutes on a single GPU suggest that this approach is accessible to organizations operating under real-world resource constraints, lowering the barrier for responsible AI deployment in industrial safety environments.

Overall, this work establishes a practical framework for integrating LLM-based assistants into constrained, real-world emergency systems. By combining quantization-aware fine-tuning, multi-metric evaluation, and safety-

oriented design principles, the study contributes to advancing reliable edge AI solutions. Future research may explore continual learning for protocol updates, multimodal integration, multi-turn interaction optimization, and controlled user studies in simulated emergency environments to further enhance robustness and usability.

Acknowledgments

This research was supported by Butkon Elevator Industry Trade Inc., Research and Development Center. Also, the authors thank Hugging Face for open-source tools [22], the bitsandbytes library[23] for quantization, and elevator safety professionals whose validation contributed to dataset quality.

Declaration of Ethical Standards

This research does not involve direct interaction with human participants or animals. Accordingly, ethical approval and informed consent were not required.

Credit Authorship Contribution Statement

All authors contributed substantially and equally to this study. The processes of conceptualization, methodology development, software implementation, validation, formal analysis, investigation, data curation, manuscript drafting, review and editing, as well as visualization were carried out collaboratively. All authors have reviewed and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Funding / Acknowledgements

This research received no external funding.

Data Availability

The dataset used in this study was constructed by the authors by synthesizing publicly available safety manuals, international elevator standards (EN 81, ASME A17.1), and academic literature. The structured question-answer pairs derived from these sources are available from the corresponding author upon reasonable request.

References

- [1] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [2] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877-1901, 2020.
- [3] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27730-27744, 2022.
- [4] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
- [5] G. Team *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.
- [6] Z. Ji *et al.*, "Survey of hallucination in natural language

- generation," *ACM computing surveys*, vol. 55, no. 12, pp. 1-38, 2023.
- [7] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 8440-8451.
- [8] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459-9474, 2020.
- [9] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *ICLR*, vol. 1, no. 2, p. 3, 2022.
- [10] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems*, vol. 36, pp. 10088-10115, 2023.
- [11] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234-1240, 2020.
- [12] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," *arXiv preprint arXiv:2010.02559*, 2020.
- [13] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.
- [14] E. Alsentzer *et al.*, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd clinical natural language processing workshop*, 2019, pp. 72-78.
- [15] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*, 2019.
- [16] P. Zare, "Enhancing maintenance strategies for elevators through fine-tuning large language models," 2024.
- [17] K. Chakma, "Evaluating enterprise product recommendation chatbot using LLM: the case of easy selection," 2025.
- [18] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74-81.
- [19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311-318.
- [20] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65-72.
- [21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.
- [22] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38-45.
- [23] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer, "8-bit optimizers via block-wise quantization," *arXiv preprint arXiv:2110.02861*, 2021.