Proceedings of the International Conference on Industrial Engineering and Operations Management
Manila, Philippines, March 7-9, 2023

Detection of Lung Cancer with Enhanced Feed Forward Backpropagation Artificial Neural Networks

Volkan Çetin, Hacı Hüsnü Yumrukaya, Çiğdem Bakır
MSc of Student Computer Enginering
MSc of Student Computer Enginering
Assistant Professor Doctor of Computer Enginering
Computer Engineering
Kutahya Dumlupinar University
Kutahya, Turkey
cigdem.bakir@dpu.edu.tr

Abstract

Although cancer is a broad term, it is an important problem that causes a high rate of death. It indicates disease that occurs when cellular changes cause cells to grow and divide uncontrollably. Cancerous cells can form tumors, damage the immune system, and cause other aberrations that prevent the body from functioning properly. In particular, lung cancer is the type of cancer with the highest death rate in the last five years. The important thing in lung cancer is early diagnosis and diagnosis to ensure the survival of patients. In this study, a model that automatically detects lung cancer has been developed in order to detect lung cancer early and take the necessary precautions at the first stage of the disease. The aim of our study is to diagnose the presence of lung cancer cells based on attributes and information from human symptoms. In our study, a dataset consisting of 26 features and 3 classes determining lung cancer as low, medium and high was used. Some preprocessing and transformations have been done to make the data more suitable for predictive analysis. The first 26 features were used as the inputs of the model. The lung cancer feature was used as the predicted output based on the input features. The values of the attributes are normalized. In the first stage, data preprocessing was carried out with various data mining techniques in order to process the data more easily and increase the performance. In the proposed model, feedforward backpropagation Artificial Neural Network model was used to detect the presence of lung cancer in the body of the person. Unlike other studies, the factors that cause lung cancer by producing our own neural network have been determined. When the results obtained are compared with the studies in the literature, it has been observed that it gives very successful results.

Keywords
lung cancer, neural network, backpropagation, artificial intelligence

1. Introduction

Lung cancer is the most important cause of death in humans. Since the symptoms of lung cancer occur in advanced stages, it leads to a high mortality rate among other types of cancer. Early diagnosis is therefore very important. Most of the lung diagnosis methods are costly and laborious. Lung cancer is a cost-effective disease that requires early diagnosis (Brunetti et al. (2022)).

Lung cancer is one of the leading types of cancer that ends in death worldwide. In the detection of lung cancer, a large number of chest radiographs are examined by expert radiologists (Joshua et al. (2020)). As a result of these examinations, misdiagnosis or differences of opinion among specialist radiologists on diagnosis and diagnoses may occur (Radhika et al. (2019)).

In our study, using the backpropagation algorithm of artificial neural networks on the lung cancer dataset, multiple predictions are carried out on patients with a rapid and accurate diagnosis on patients with suspected lung cancer. Thanks to this diagnosis, the patient's risk of developing cancer is estimated as low, medium and high. In our study, it was aimed to develop an automatic model with developed artificial neural networks and lung data. Feature extraction was performed with the model we created, and then it was classified and compared with different machine learning methods. Success rates were calculated according to various performance criteria that were compared with the results.

This classification has contributed positively to its success. In addition, the difference of our study from other studies is that we create our own artificial neural network. We will improve the study by performing this neural network on different datasets with different parameters.

The article is organized as follows: In the 2nd part of our study, the studies carried out in the literature, in the 3rd part the method, in the 4th part the proposed method, in the 5th part the application and in the last part the results we obtained as a result of the study are given.

1.1 Objectives

The stages of our work are given below:
• Acquisition of the lung cancer dataset.
• Preprocessing of data using data mining techniques.
• Determining the Model
• Training of the obtained data set.
• Performing performance analyzes on the detection of cancer risk.
In the training dataset for the lung data we used, the 5-fold layered crossover validation was used to evaluate the performance of the models. In addition, after choosing the best model according to k-fold cross validation, all training data were trained and next evaluated on the test dataset. Logistic activation to backpropagation algorithm function was used and the sum of squared errors cost function was calculated. We trained our model using the standard back propagation algorithm.

2. Literature Review

Lung cancer is one of the malignant tumors with the highest mortality in the world. The overall five-year survival rate of lung cancer is relatively lower than many leading cancers. Early diagnosis and prognosis of lung cancer are essential to improve the patient's survival rate (Huang et al. 2023) In the literature, studies conducted in the last 5 years with the detection of lung cancer are given.

Lemieux et.al, proposed detection of early-stage lung cancer in sputum using automated fow cytometry (Lemieux et al. (2023)). This study prepared single cell suspensions from induced sputum samples. Moreover, the dataset they used were labeled with a viability dye to exclude dead cells, antibodies to distinguish cell types, and a porphyrin to label cancer-associated cells. With machine learning algorithm was developed to distinguish cancer from non-cancer samples from 150 patients at high risk of whom 28 had lung cancer. However, this study does not provide definitive results in detecting lung cancer in sensitive data.

Dritsas et al. used machine learning (ML) methods on various datasets tried to identify individuals at high risk for developing lung cancer (Dritsas, and Trigka (2022)). The success of the study was evaluated by well-known measures such as precision, recall, F-Measure, accuracy and area under the curve (AUC). Various machine learning models, including NB, BayesNet, SGD, SVM, LR, ANN, KNN, J48, LMT, RF, RT, RepTree, RotF, and AdaBoostM1, were evaluated in terms of accuracy, precision, recall, F-Measure and AUC. However, the detection of cancerous cells in the study could not be determined in a long time.

Alsinglawi et al. introduces a predictive Length of Stay (LOS) framework for lung cancer patients using machine learning (ML) models (Alsinglawi et al. (2022)). ML methods were used to estimate the length of stay of patients hospitalized in the intensive care unit due to lung cancer. In the study, The Random Forest (RF) Model outperformed the other models. In order to make prediction-based classification, first of all, the problem of unbalanced classification was emphasized. They used different evaluation metrics (Index Balanced Accuracy (IBA), Geometric Mean Score (GMS), Precision, Sensitivity, Specifcity, F1-score) to evaluate the predictive models' performance.

Wang et.al, identified specifc biomarkers for lung cancer (Wang et al. 2022). They obtained three gene-expression profles and screened for differentially expressed genes (DEGs) between lung cancer and normal lung tissue. Machine learning methods were used to identify the optimal diagnostic biomarkers for lung cancer such as logistic regression, and support vector machine recursive feature elimination.

Yang et al. (2022), integrated genomic, clinical and demographic data of lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) patients from The Cancer Genome Atlas (TCGA) and introduce copy number

variation (CNV) and mutation information of 15 selected genes to generate predictive models for recurrence and survivability. We compare the accuracy and benefits of three well-established machine learning algorithms: decision tree methods, neural networks and support vector machines.

Ishii et al. (2022), developed gene alteration prediction model for primary lung cancer. They developed machine learning models to predict several common gene alterations in primary lung cancer based on digital images of cytology specimens. This model is a promising tool for precision medicine to screen candidate gene alterations and provide beneficial information for lung cancer patients. Boddu et al. (2022), identified the impact of machine learning and artificial intelligence in the covid 19 outbreak for lung cancer. Machine learning techniques show excellent precision for distinguishing COVID-19 from non-COVID-19 chest pneumonia. These methods have made it easier to assess these pictures automatically.

Gould et al. (2021), used machine learning for early lung cancer identification using routine clinical and laboratory data . They developed model predict future diagnısis of lung cancer basis of outlne clinical and laboratory data by using machine learning. Patra et al. (2020), analyzed various machine learning classifiers techniques to classify available lung cancer data in UCI machine learning repository in to benign and malignant. The input data is prepossessed and converted in to binary form followed by use of some well known classifier technique in Weka tool to classify the data set in to cancerous and non cancerous.

3. Methods

Artificial neural networks work similarly to the learning and thinking processes of the human brain.
It is a machine learning model in which mathematical models are used. Artificial neural networks, one or may have several operational layers between multiple input and output layers, and this between the layers there is a structure created by weights and integrators. Inputs represent data coming from the external environment to the network. The outputs represent the information learned by the network. it does. After the information is transmitted from the input layer to the network, the weight values of the network in the operational layer are sent to the output layer. If there is more than one computational neural network multi-layer artificial neural network if there are layers and neurons, if it consists of a single layer, single-layer It is called an artificial neural network. Artificial neural networks can consist of various design patterns and It can be trained using learning algorithms (Muhammad et al. 2019)).

Artificial neural networks have various uses and can be trained using different learning algorithms. It can be used in Classification, Regression, Clustering and other machine learning tasks (Desai and Shah (2021)). They can also be used in fields such as artificial neural networks, voice and image recognition, commerce, health, military and environment. Examples such as brain modeling studies, cancer cell detection, classification of blood samples can be given for the health sector. Artificial neural networks can be divided into two in terms of their advantages and disadvantages.

Artificial Neural Networks Advantages

Learning: Artificial neural networks have the ability to learn over data. Structure of the data and can make inferences based on these data. learning, artificial neural. It allows networks to process data better (Yan et al. (2020)).

Learning Ability: With the ability to learn artificial neural networks, the model given to the input
can produce output accordingly.
Learning Rate: Artificial neural networks are better than other artificial intelligence techniques in terms of learning speed is faster than. This allows the artificial neural network model to learn faster and ensures good performance.
Learning Capacity: Artificial neural networks have the ability to learn on very large data sets. This means that the neural network model is better based on more data and provides performance.
Learning Depth: Artificial neural networks integrated with deep learning techniques
gains a deep learning ability.
Learning Flexibility: Artificial neural networks have a flexible learning ability. The artificial neural network model has the ability to learn on different data sets and It allows to produce different outputs.

Artificial Neural Networks Disadvantages

Requires Training: Neural networks require training on datasets. This is artificial It requires the time and effort required for the neural network model to acquire the learning ability (Shakeel et al. (2019)).
Limited Learning Capacity: Artificial neural networks only work on given datasets has the ability to learn. For this reason, the learning ability of the artificial neural network model
associated with the datasets used.

Learning Flexibility Limited: Artificial neural networks have the ability to learn flexibly. But flexible learning ability is limited.
Learning Performance is Limited: Artificial neural networks are better than other networks in terms of learning performance lower than artificial intelligence techniques.
Learning Process Is Time-consuming: Neural networks are better than other artificial neural networks in terms of learning process longer than intelligence techniques.

Backpropagation Definition*:* The backpropagation algorithm was developed in the mid-eighties. It is one of the learning algorithms used for Neural Networks. To put it briefly; in the first part, random numbers are assigned to the neuron weights. Then network training starts. Weights are calculated from input layers to hidden layers and results are found using an Active Function. Using these results, weights are calculated from hidden layers to output ayer and active function and output layer values are obtained again. In the last part, changes in weights are calculated and updated backwards, that is, from output Layer to hidden layers, from hidden layers to input layers.

The backpropagation algorithm is given in Figure 1. It searches for the minimum value of the error function in the weight domain using a technique called the delta rule or gradient descent (Nasien et al. (2022)). It is thought that the weights that minimize the error function are a solution to the learning problem.
Backpropagation is the core of Neural Network training. Loss value obtained in epoch It is a method of adjusting the weights of a neural network based on weights correctly adjusting it allows us to reduce the error rate and increase its generalization, making the model reliable allows us to bring .

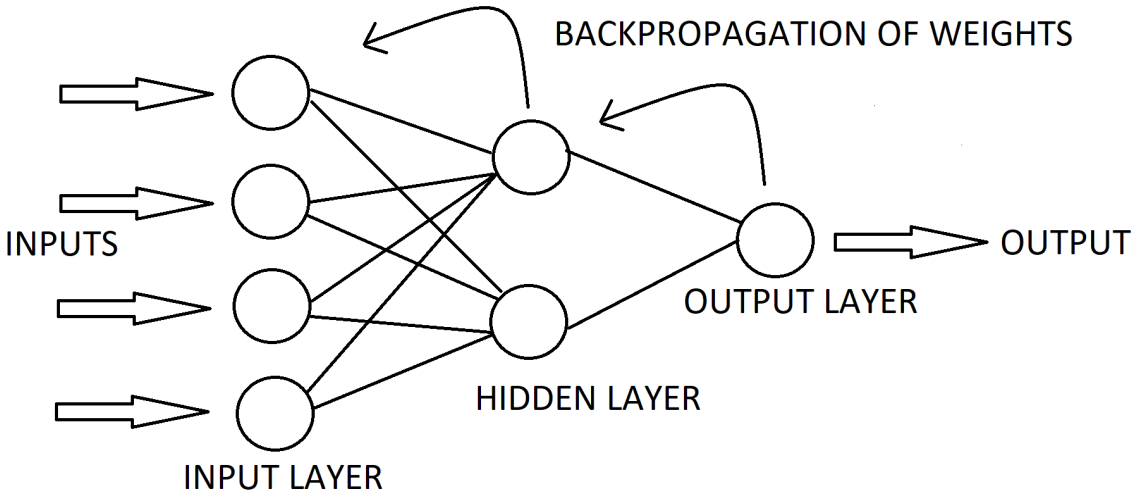

Figure 1. Backpropagation algorithm

4. Data Collection

In our study, lung data from https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link?resource=download was used to predict and detect lung cancer . The dataset consists of 26 features and 3 classes. The characteristics of this dataset are given below:

Characteristics of the dataset (26 pieces): Ranking Index, Unique Identification Index, Age, Gender, Air Pollution, Alcohol Use, Dust Allergy, Occupational Hazards, Genetic Risks, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoking, Chest As a result of features such as Pain, Mouth Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Difficulty in Swallowing, Numbness in Fingers, Body Cold, Dry Cough, Snoring, lung cancer predicted levels are determined as low, medium and high levels. In this way, the lung cancer status of individuals can be detected before the disease progresses.

Classification Method to be Used: The flow chart of the model we proposed is shown in Figure 2. Our study was created in two stages. In the first stage, the data were normalized and categorized. In the second stage, the risk of developing lung cancer was classified by using a supervised learning algorithm for Backpropagation, Multilayer Perceptron (Artificial Neural Networks) training. One Step Secant Backpropagation algorithm is used in this classification method. The difference of our study from other studies is that we created our own neural network. We create this neural network automatically on different datasets and we can apply it to each dataset. We also conducted various performance analyzes on the risk of developing cancer. The results are quite promising (Kasim et al. (2022)).

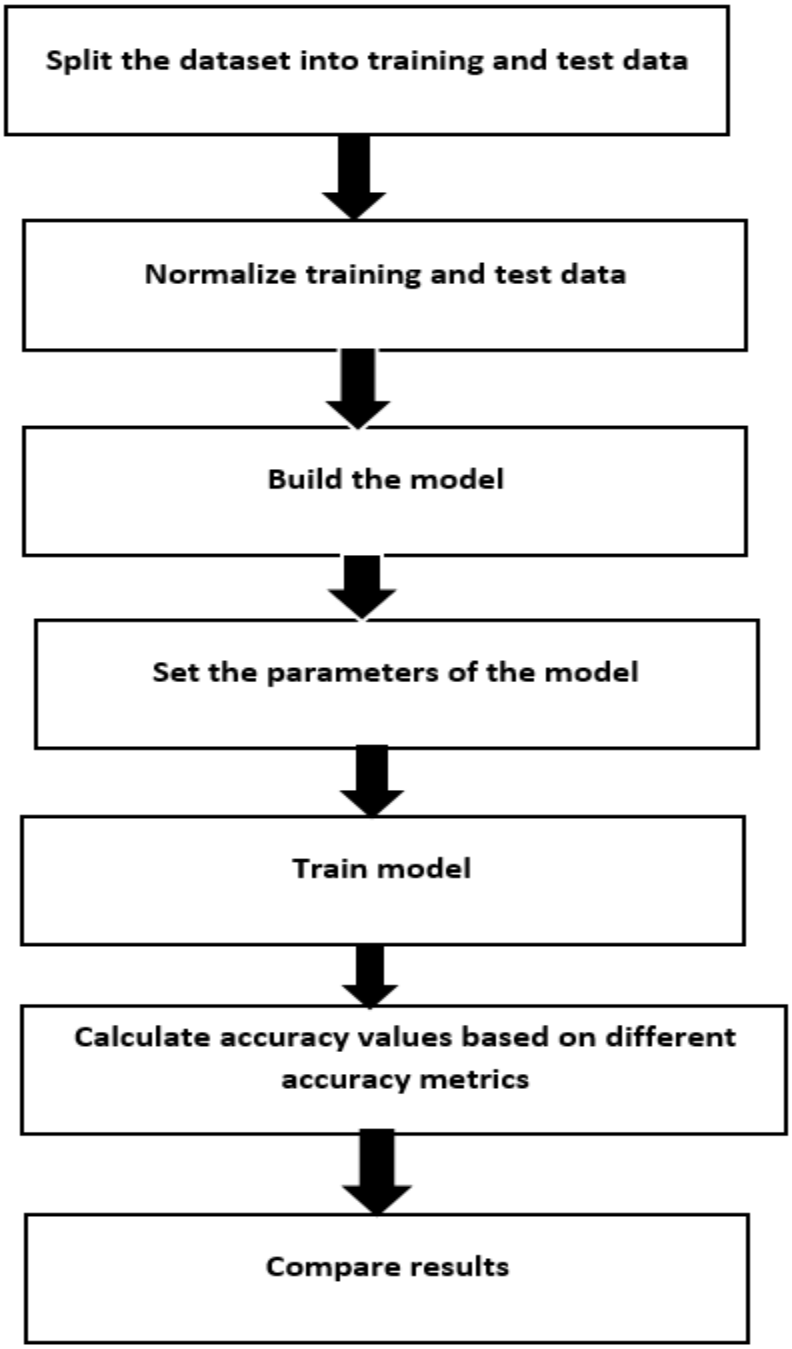

Figure 2. Flow chart of the proposed model

5. Results and Discussion

F1 Score, Recall and Precision were used as model evaluation criteria (Banerjee and Das (2020)). Accuracy results are given according to all metrics used in Table 1. Begins in the proposed model for the classification of lung cancer low, medium and high risk groups artificial neural network. And comparing the accuracy with the proposed model, recall while accuracy can be seen to increase was less.

Recall: It measures the proportion of actual positive that are correctly identified.

$$P_r = \frac{T_{pos} + F_{neg}}{T_{pos}}$$

Precision: It measure the proposition of positive identification is actually correct.

$$P_{prec} = \frac{T_{pos}}{T_{pos} + F_{pos}}$$

F1 Score: F1 score is the average of both precision and recall.

$$P_{F1} = 2 * \frac{Recall * Precision}{Recall + Precision}$$

Table 1. Results of the proposed model according to various metrics

Metrics	Accuracy	Precision	Recall	F1 Score
Proposed Model	98,75	95,10	60,42	72,60

6. Conclusion

Lung cancer is the formation of tumors in the lung as a result of abnormal and uncontrollable proliferation of lung cells. Lung cancer has become the most common type of cancer in the world in parallel with the increase in smoking habits. Due to the importance of the subject, November has been chosen as 'Lung Cancer Awareness Month' around the world. Although it is a fatal disease, early diagnosis of the disease is very important. As in all types of cancer, in the treatment of lung cancer, it is possible to diagnose and fully treat the disease at the first stage with early diagnosis. Identifying the disease in its early stages is very important in terms of treatment.

In our study, this important disease was handled and the early diagnosis and diagnosis of the disease was emphasized. An automated neural network model has been developed to detect the disease early. In our study, a data set containing information about the risk status of people with and without lung cancer (3 classes as low, medium and high) was used. Some preprocessing and transformations have been made on the utilized datasets in order to make the data more suitable for predictive analysis. With this research, the presence of lung cancer cells can be detected based on the attributes and information from human symptoms using Artificial Neural Network. Then, the Artificial Neural Networks model was trained and validated. In the Artificial Neural Networks model, it was able to predict the presence of lung cancer with 98.75% accuracy after 10000 learning cycles with less than 1% training error rate. In addition, the model obtained showed that chronic diseases were the most influential feature on the presence of lung cancer. This study showed that the neural network can diagnose lung cancer, so it can be used by doctors as a diagnostic tool. This study showed that the neural network can diagnose lung cancer, so it can be used by doctors as a diagnostic tool.
In the future, it is aimed to increase the success of the study by automatically modeling different network structures on different and very large real datasets. In addition, the difference of our study from other studies is that we create

our own artificial neural network. We will improve the study by performing this neural network on different datasets with different parameters.

References

Alsinglawi, B., Alshari, O., Alorjani, M., Mubin, O., Alnajjar, F., Novoa, M., & Darwish, O. , An explainable machine learning framework for lung cancer hospital length of stay prediction. *Scientific reports*, *12*(1), 1-10, 2022.

Banerjee, N., & Das, S. (2020, March). Prediction lung cancer–in machine learning perspective. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-5). IEEE.

Boddu, R. S. K., Karmakar, P., Bhaumik, A., Nassa, V. K., & Bhattacharya, S., Analyzing the impact of machine learning and artificial intelligence and its effect on management of lung cancer detection in covid-19 pandemic. *Materials Today: Proceedings*, *56*, 2213-2216,2022.

Brunetti, A., Altini, N., Buongiorno, D., Garolla, E., Corallo, F., Gravina, M., ... & Prencipe, B. , A Machine Learning and Radiomics Approach in Lung Cancer for Predicting Histological Subtype. *Applied Sciences*, *12*(12), 5829, 2022.

Desai, M., & Shah, M. , An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). *Clinical eHealth*, *4*, 1-11, 2021.

Dritsas, E., & Trigka, M. , Lung Cancer Risk Prediction with Machine Learning Models. *Big Data and Cognitive Computing*, *6*(4), 139, 2022.

Gould, M. K., Huang, B. Z., Tammemagi, M. C., Kinar, Y., & Shiff, R., Machine learning for early lung cancer identification using routine clinical and laboratory data. *American Journal of Respiratory and Critical Care Medicine*, *204*(4), 445-453, 2021.

Huang, S., Yang, J., Shen, N., Xu, Q., & Zhao, Q. (2023, January). , Artificial intelligence in lung cancer diagnosis and prognosis: current application and future perspective. In *Seminars in Cancer Biology*. Academic Press.

Ishii, S., Takamatsu, M., Ninomiya, H., Inamura, K., Horai, T., Iyoda, A., ... & Takeuchi, K., Machine learning-based gene alteration prediction model for primary lung cancer using cytologic images. *Cancer Cytopathology*, *130*(10), 812-823, 2022.

Joshua, E. S. N., Chakkravarthy, M., & Bhattacharyya, D., An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study. *Rev. d'Intelligence Artif.*, *34*(3), 351-359, 2020.

Lemieux, M. E., Reveles, X. T., Rebeles, J., Bederka, L. H., Araujo, P. R., Sanchez, J. R., ... & Rebel, V. I. , Detection of early-stage lung cancer in sputum using automated flow cytometry and machine learning. *Respiratory Research*, *24*(1), 1-16, 2023.

Mohamed Kasim, J., & Murugan, B., Lung Cancer Segmentation Using Enriched Fuzzy Back-Propagation Neural Network. In *Cybernetics Perspectives in Systems: Proceedings of 11th Computer Science On-line Conference 2022, Vol. 3* (pp. 502-518), 2022. Cham: Springer International Publishing.

Muhammad, W., Hart, G. R., Nartowt, B., Farrell, J. J., Johung, K., Liang, Y., & Deng, J. (2019). Pancreatic cancer prediction through an artificial neural network. *Frontiers in Artificial Intelligence*, *2*, 2.

Nasien, D., Enjeslina, V., Adiya, M. H., & Baharum, Z. (2022, August)., Breast Cancer Prediction Using Artificial Neural Networks Back Propagation Method. In *Journal of Physics: Conference Series* (Vol. 2319, No. 1, p. 012025). IOP Publishing.

Shakeel, P. M., Burhanuddin, M. A., & Desa, M. I., Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks. *Measurement*, *145*, 702-712,2019.

Patra, R., Prediction of lung cancer using machine learning classifier. In *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1* (pp. 132-142), 2020.. Springer Singapore.

Radhika, P. R., Nair, R. A., & Veena, G. (2019, February). A comparative study of lung cancer detection using machine learning algorithms. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-4). IEEE.

Wang, F., Su, Q., & Li, C., Identidication of novel biomarkers in non-small cell lung cancer using machine learning. *Scientific Reports*, *12*(1), 16693, 2019.

Yan, R., Ren, F., Wang, Z., Wang, L., Zhang, T., Liu, Y., ... & Zhang, F., Breast cancer histopathological image classification using a hybrid deep neural network. *Methods*, *173*, 52-60, 2020.

Yang, Y., Xu, L., Sun, L., Zhang, P., & Farid, S. S. (2022). Machine learning application in personalised lung cancer recurrence and survivability prediction. *Computational and Structural Biotechnology Journal*, *20*, 1811-1820, 2022.

Biography

Volkan Çetin is from Eskişehir. I graduated from Hoca Ahmet Yesevi University, Management Information Systems. I am doing my master's degree in computer engineering at Kütahya Dumlupınar University. I have studies in the fields of 3D printers and image processing.

Hacı Hüsnü Yumrukaya.is from Eskişehir. I graduated from Kütahya Dumlupınar University Computer Engineering Department and completed the department with the first place. I am currently doing my master's degree with thesis in Kütahya Dumlupınar University, computer engineering department. I did image processing and robotic surface studies.

Çiğdem Bakır is a Assistant Professor of Software Engineering at the Engineering Faculty – Kutahya University, Turkey. She received the B.S. degree in computer engineering from the University of Sakarya, in 2010, and the M.S. and Ph.D. degrees in computer engineering from Yildiz Technical University, Istanbul. She is currently pursuing the doctorate degree in computer science with the University of Yildiz Technical, Istanbul. She was a Research Assistant at Yildiz Technical University and Igdir University. She was an Instructor at Erzincan Binali Yildirim, from 2020 to 2021. She has been an Assistant Professor with the Software Engineering Department, Dumlupinar University, since 2021. Her research interests include information security, distributed database, big data, blockchain technology, cloud computing, and computer networks.