DOI: https://doi.org/10.33103/uot.ijccce.22.1.6

Lung Cancer Prediction and Risk Factors Identification using Artificial Neural Network

Israa N. Mahmood[1], Hasanen S. Abdullah[2]

[1,2]Department of Computer Science, University of Technology, Baghdad, Iraq

[1]cs.19.78@grad.uotechnology.edu.iq, [2]110014@uotechnology.edu.iq

Abstract*— Lung cancer is one of the most fatal cancers in the world for both genders. It has a high mortality rate compared to other types of cancer. Early detection can save lives and enhance the treatment process. As a result, the demand for approaches to detect cancer at an early stage is growing. In this paper, an Artificial Neural Network (ANN) model is developed to identify the level of having lung cancer based on environmental, diagnostic, and statistical factors. The features that highly affect the risk level of lung cancer were identified. The model's performance was assessed using a variety of criteria, including accuracy, precision, recall, and f-measure. Experimental results show that the model attains a high accuracy rate of 91.79% and risk factors like obesity, alcohol use, genetic risk, and coughing of blood can lead to lung cancer.*

Index Terms*— Classification, Lung Cancer, Machine Learning Algorithm, Artificial Neural Network***.**

I. INTRODUCTION

Lung cancer is one of the most serious types of cancer, where the death rate surpasses the breast, colon, and prostate cancer combined. In 2020, 2.2 million were diagnosed of lung cancer and the mortality rate exceeded 1.8 million, which constitute 18% of total cancer death [1]. Most lung cancer patients are diagnosed in a late stage, where treatment becomes less efficient. The success of the treatment program depends on the early detection of cancer.

In recent years, several industries such as telecommunication [2], financial services [3], disaster prediction [4], and medical diagnosis [5] use machine-learning models to enhance their performance. These models can easily identify hidden patterns of data in different fields without human intervention. The public health systems are rich in data but deficient in knowledge. The system creates massive amounts of data but it lacks the necessary analysis capabilities to educe knowledge. Machine learning approaches can be used to help experts in diagnosing diseases and extract knowledge.

Traditional classification algorithms can be used to recognize people at a high-risk level of having lung cancer. Identifying such patients can lead to early detection of cancer, which results in a higher survival rate.

The primary contribution of this research is to develop a neural network model that can predict the risk level of having lung cancer based on statistical (age, gender), diagnostic (smoking, alcohol usage), and Environmental (air pollution) characteristics and identify which features are highly related to the risk level. Building a model with high accuracy can result in saving the patient’s life.

Received 22/June/2021; Accepted 26/September/2021

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.1.6

The remainder of this research is structured as follows; different classification algorithms to diagnose cancer are discussed in section 2. The proposed model structure and the used data set are presented in section 3. The results and discussion is illustrated in section 4. The conclusion of this study is discussed in section 5.

II. RELATED WORK

During the last few years, there has been a growth in the studies that use classification algorithms to diagnose different kinds of cancer such as breast [6], brain [7], prostate [8], lung [9], and cervical [10]. Several studies will be explored in this section regarding cancer detection using machine learning algorithms.

Asri et al. [11] compared the performance of different classification algorithms in diagnosing breast cancer using Wisconsin breast cancer dataset. The result shows that Support Vector Machine (SVM) achieves a higher accuracy rate than K-Nearest Neighbor (KNN), Naïve Bayes, and C4.5 Decision Tree.

Amrane et al. [12] applied two classification algorithms; Naïve Bayes and KNN to identify malign breast cancer using Wisconsin breast cancer dataset. KNN attains a better accuracy rate than Naïve Bayes classifier. Polly et al. [13] proposed a computerized system to distinguish between normal and abnormal tumors in the brain and to specify the type of the abnormal tumor. The system used k-means algorithm for segmentation and SVM to classify the tumor type.

Wang et al. [14] developed predictive models to diagnose prostate cancer. The proposed model used various learning models such as Least Square SVM, Random Forrest, SVM, and Artificial Neural Network (ANN). The ANN outperforms other models and achieved a 95% accuracy rate in detecting significant prostate cancer.

Murugan et al. [15] analyzed the behavior of different classification algorithms such as Random Forrest, KNN, and SVM to detect skin cancer. The implemented model used watershed method for segmentation. Features were extracted from the resulted segments and used for classification.

Faisal et al. [16] suggested a majority voting framework to diagnose lung cancer. The proposed framework selects the top 3 classifiers among a list of traditional classifiers like SVM, Neural Network, Multi-Layer Perceptron, Naïve Bayes, and C4.5 Decision Tree. The result shows that the majority voting method achieved an 88% accuracy rate.

Das et al. [17] proposed an intelligent system to automatically detect liver cancer. The model used watershed and Gaussian methods to extract cancer lesion from Computed Tomography (CT) scan images and used a deep learning model for classification. The model achieved a high accuracy rate of 99.3%.

Shivaprasad and Naveena [18] used two classification algorithms to check their effectiveness in diagnosing lung cancer. It was observed that Logistic Regression achieves better results than KNN when the number of iterations increases.

Several methods were used to detect different types of cancer. All of these studies focus on the advantage of early detection of cancer that can result in a high recovery rate. Our research focuses on identifying people with a high risk of having lung cancer using ANN model.

Received 22/June/2021; Accepted 26/September/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.1.6

III. PROPOSED SYSTEM ARCHITECTURE

A. Data Set

In this research, the lung cancer dataset that can be found in Kaggle were used [19]. The dataset predicts the level of having lung cancer according to different types of features:

- **Statistical Features:** age and gender.
- **Diagnostic Features:** alcohol use, balanced diet, chest pain, clubbing of fingernails, coughing of blood, chronic lung disease, dry cough, dust allergy, fatigue, frequent cold, genetic risk, obesity, passive smoking, shortness of breath, smoking, snoring, swallowing difficulty, weight loss, and wheezing.
- **Environmental Features:** air pollution and occupational hazard.

The data set consists of 1000 instances with 3 types of risk levels (low, medium, high). However, this study only uses low and high risk levels. So the number of instances is 668, and the distribution of the instances according to the level is illustrated in *Fig. 1*.

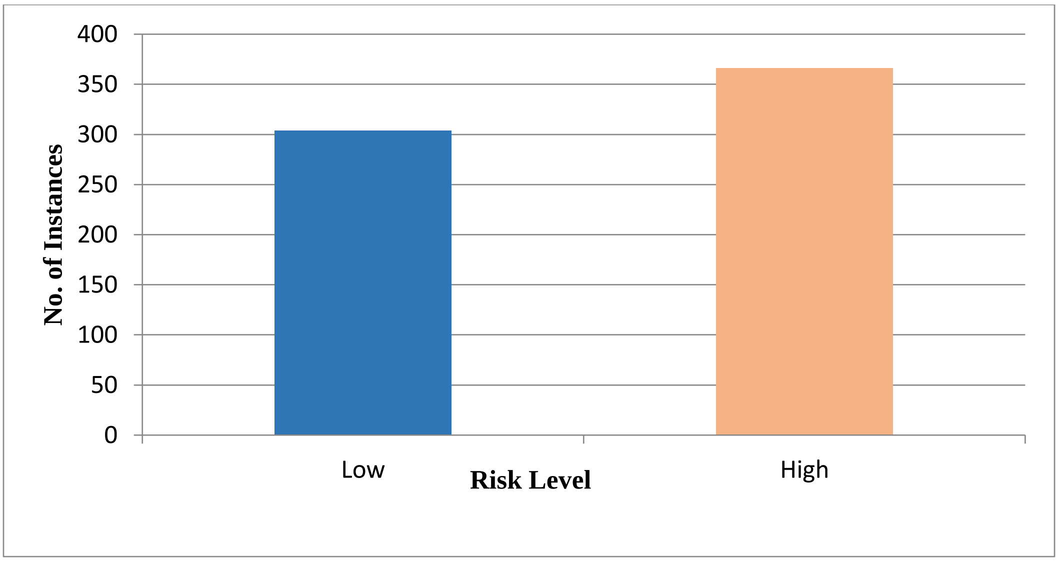

FIG. 1. DISTRIBUTION OF INSTANCES BASED ON RISK LEVEL

B. Proposed System Architecture

ANN is designed to mimic the behavior of the human brain. It consists of input, hidden, and output layers. The input layer deals with data such as image, text, or audio. While the hidden layer analyzes and processes the data and the output layer produces one or more output. The main element in ANN is the neuron, which accepts several inputs, multiplies them by weights, and passes the sum of multiplication to the activation function. The output is then slipped to one or more neurons. In our research, an ANN model that consists of 3 hidden layers and 1 output layer were developed as shown in *Fig. 2*. The algorithm of the proposed system is illustrated in Algorithm 1. The architecture that achieved a train accuracy rate higher than the threshold value was selected. The model was implemented using python language.

Received 22/June/2021; Accepted 26/September/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.1.6

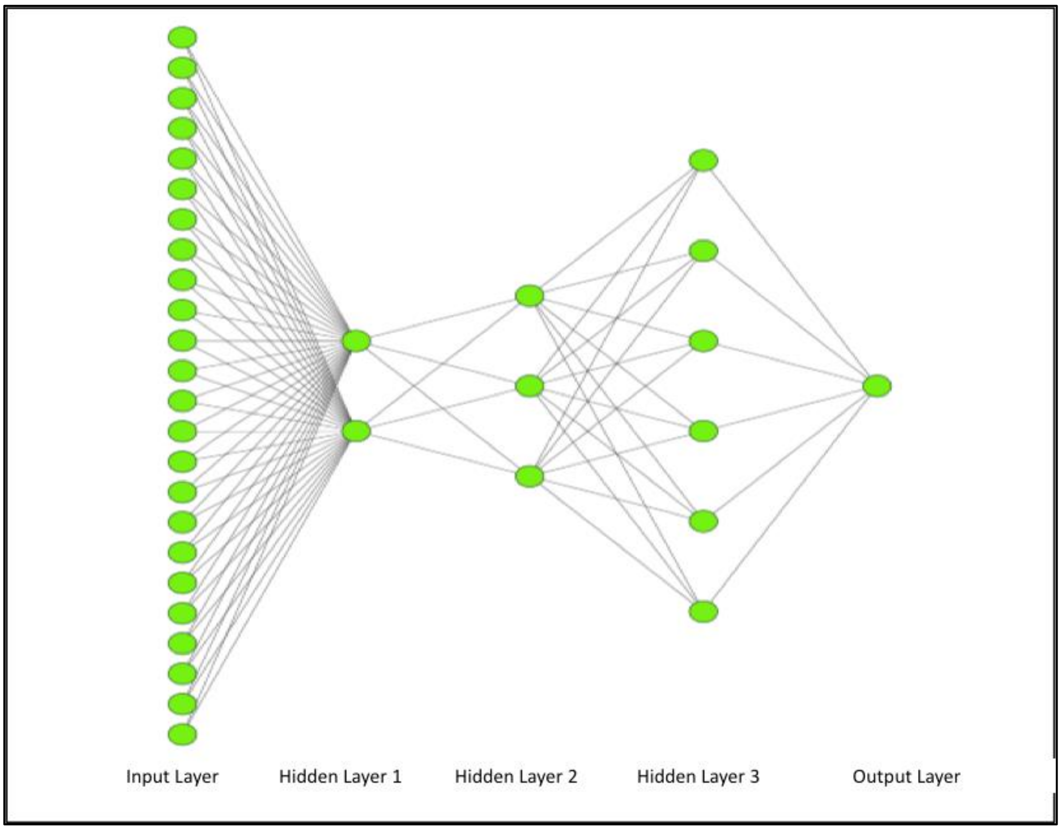

FIG. 2. ARCHITECTURE OF THE PROPOSED ARTIFICIAL NEURAL NETWORK TO PREDICT LUNG CANCER

```
Algorithm 1:  Lung cancer prediction model algorithm
Input:
      $R$: training data set
      $N_e$: Number of epochs.
      $T$: Accuracy Threshold
Output: Artificial Neural Network to predict lung cancer risk level $M$
     Perform data pre-processing (data cleaning, remove outlier, remove unique features…etc.)
     $F$= False.
     Repeat
          Create a $NN$ neural network with $H$ hidden layers and $N$ neurons at each layer.
          Set weights $W$ and bias $b$ randomly in $NN$
          For $j < N_e$ do
             Feed $R$ to $NN$.
             Update weights and bias in $NN$ using back propagation.
          End For
          If $\text{Accuracy}_{train} >= T$
             $F$ = True
          Else
             Update number of hidden layers $H$.
             Update number of neurons $N$ at each layer.
          End if
     Until ($F$ = True)
Return $M$
```

Received 22/June/2021; Accepted 26/September/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.1.6

IV. RESULTS AND DISCUSSION

A. Evaluation Measures

Several metrics are used in traditional classification algorithms to evaluate their performance such as accuracy, recall, f-measure, and precision. These measures are calculated from the confusion matrix elements [20]. It consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

- Accuracy measures the rate of instances that were correctly classified as shown in (1)

$$Accuracy = \frac{TP + TN}{P + N} \qquad (1)$$

- Recall calculates the rate of TP compared to P as illustrated in (2)

$$Recall = \frac{TP}{P} \qquad (2)$$

- Precision calculates the rate of TP compared to P’ as depicted in (3)

$$Precision = \frac{TP}{P'} \qquad (3)$$

- F-measure merges recall and precision measures into one measure in a consistent mean as presented in (4).

$$F - measure \; = \; 2 * \frac{precision * recall}{precision \; + recall} \qquad (4)$$

B. Analysis and Discussion of Lung Cancer Prediction Model

The performance of the proposed model was compared against several classification algorithms such as Gaussian NB, J48, SVM, and Decision Tree algorithms to prove the effectiveness of the model. The ANN model surpasses other models and achieves a high accuracy rate of 91.79%. Gaussian NB and Decision Tree were able to attain a good accuracy rate of 90% and 88.75% accordingly. However, SVM and J48 have the lower accuracy rate of 65% and 83% respectively. Table I demonstrates that the ANN model was also able to achieve better results in precision, recall, and F-measure.

TABLE I. PERFORMANCE MEASUREMENT FOR DIFFERENT CLASSIFICATION ALGORITHMS

Model	Accuracy %	Recall %	Precision %	F-Measure %
ANN	91.79	91.78	92.13	91.93
Gaussian NB	90.00	90.90	90.90	90.00
Decision Tree	88.75	88.73	88.75	88.96
J48	83.75	84.21	83.89	83.72
SVM	65.00	63.18	62.50	62.66

The ANN model also achieved the minimum error rate of ~8% compared to the other models as shown in *Fig. 3*.

Received 22/June/2021; Accepted 26/September/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.1.6

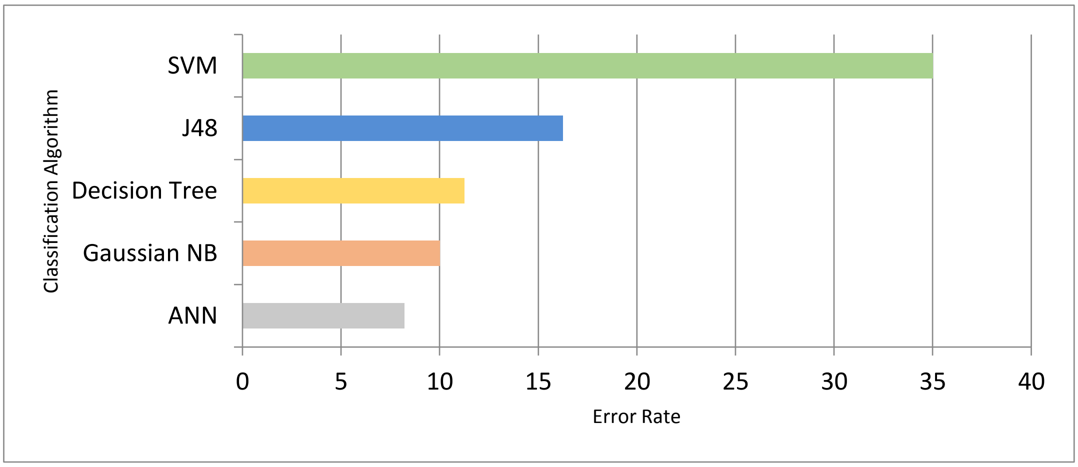

FIG. 3. ERROR RATE PERCENTAGE FOR DIFFERENT CLASSIFICATION ALGORITHMS.

C. Analysis of Risk Factors Importance

A heat map was used to identify the leading risk factors in determining the level of having lung cancer. The features heat map in *Fig. 4* shows that factors like coughing of blood, obesity, alcohol use, and genetic risk can highly affect the level of lung cancer. Identifying people with these factors can help in the early detection of lung cancer by performing a periodic examination and can result in saving their lives.

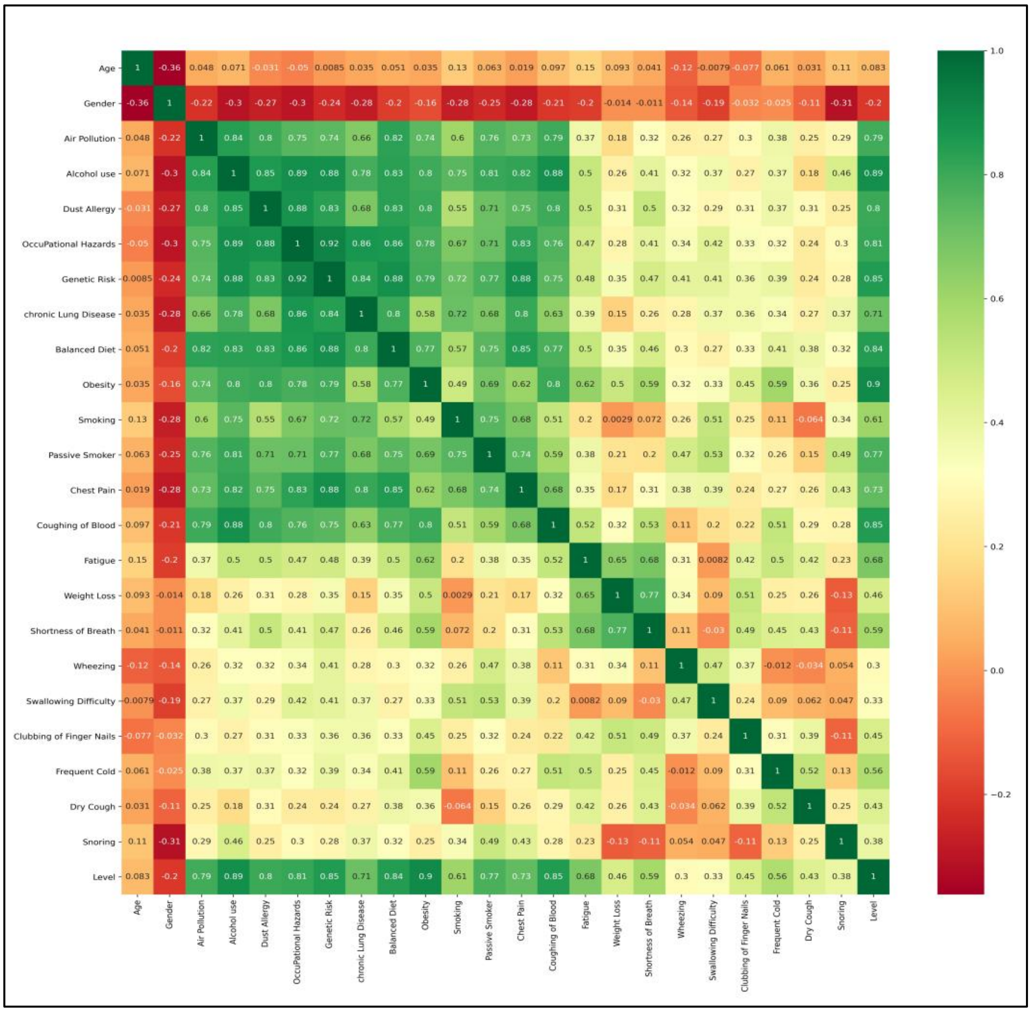

FIG. 4. HEAT MAP OF FEATURE IMPORTANCE

Received 22/June/2021; Accepted 26/September/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.1.6

V. CONCLUSIONS

Lung cancer is one of the dangerous diseases in the world. Most of its patients are diagnosed at a late stage, which make the treatment process less efficient. Early detection can save the patient's life. Recently, medical staff used machine-learning algorithms and artificial neural network to help them in analyzing the medical data. Developing an accurate classifier with high computational efficiency in the medical field is a major challenge. In this study, an ANN model was developed to detect people at high risk for lung cancer based on behavioral factors. This study also analyzed the features and identified which risk factors can lead to lung cancer. The primary purpose of this study was to warn people of high risk so that they can perform a periodic examination and detect cancer at an early stage. Experimental results illustrate the high ability of ANN to diagnose lung cancer.

REFERENCES

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[2] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0191-6.

[3] V. N. Dornadula and S. Geetha, "Credit Card Fraud Detection using Machine Learning Algorithms," *Procedia Comput. Sci.*, vol. 165, pp. 631–641, 2019, doi: https://doi.org/10.1016/j.procs.2020.01.057.

[4] C. Bi *et al.*, "Machine learning based fast multi-layer liquefaction disaster assessment," *World Wide Web*, vol. 22, no. 5, pp. 1935–1950, 2019.

[5] J. Ker, Y. Bai, H. Y. Lee, J. Rao, and L. Wang, "Automated brain histology classification using machine learning," *J. Clin. Neurosci.*, vol. 66, pp. 239–245, 2019, doi: https://doi.org/10.1016/j.jocn.2019.05.019.

[6] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in *Proceedings of the 2nd international conference on machine learning and soft computing*, 2018, pp. 5–9.

[7] A. Sawant, M. Bhandari, R. Yadav, R. Yele, and M. S. Bendale, "Brain cancer detection from mri: A machine learning approach (tensorflow)," *Brain*, vol. 5, no. 04, 2018.

[8] Y. K. Tsehay *et al.*, "Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images," in *Medical Imaging 2017: Computer-Aided Diagnosis*, 2017, vol. 10134, p. 1013405.

[9] S. Bhatia, Y. Sinha, and L. Goel, "Lung Cancer Detection: A Deep Learning Approach BT - Soft Computing for Problem Solving," 2019, pp. 699–705.

[10] A. Ghoneim, G. Muhammad, and M. S. Hossain, "Cervical cancer classification using convolutional neural networks and extreme learning machines," *Futur. Gener. Comput. Syst.*, vol. 102, pp. 643–649, 2020, doi: https://doi.org/10.1016/j.future.2019.09.015.

[11] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[12] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2018, pp. 1–4.

[13] F. P. Polly, S. K. Shil, M. A. Hossain, A. Ayman, and Y. M. Jang, "Detection and classification of HGG and LGG brain tumor using machine learning," in *2018 International Conference on Information Networking (ICOIN)*, 2018, pp. 813–817, doi: 10.1109/ICOIN.2018.8343231.

[14] G. Wang, J. Y. C. Teoh, and K. S. Choi, "Diagnosis of prostate cancer in a Chinese population by using machine learning methods," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2018-July, pp. 1–4, 2018, doi: 10.1109/EMBC.2018.8513365.

[15] A. Murugan, S. A. H. Nair, and K. P. S. Kumar, "Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers," *J. Med. Syst.*, vol. 43, no. 8, p. 269, 2019, doi: 10.1007/s10916-019-1400-8.

[16] M. I. Faisal, S. Bashir, Z. S. Khan, and F. Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer," *2018 3rd Int. Conf. Emerg. Trends Eng. Sci. Technol. ICEEST 2018*, pp. 1–4, 2019, doi: 10.1109/ICEEST.2018.8643311.

[17] A. Das, U. R. Acharya, S. S. Panda, and S. Sabut, "Deep learning based liver cancer detection using

Received 22/June/2021; Accepted 26/September/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.1.6

watershed transform and Gaussian mixture model techniques," *Cogn. Syst. Res.*, vol. 54, pp. 165–175, 2019, doi: https://doi.org/10.1016/j.cogsys.2018.12.009.

[18] P. Shivaprasad and C. Naveena, "Design and Implementation of Multi-class Logistic Regression for Effective Classification of Low, Medium and High Risk Lung Cancer Problem," *Adv. VLSI, Signal Process. Power Electron. IoT, Commun. Embed. Syst. Sel. Proc. VSPICE 2020*, vol. 752, p. 317, 2021.

[19] D. Maharana, "Kaggle Lung Cancer Dataset." https://www.kaggle.com/divyanimaharana/lung-cancer-dataset (accessed May 01, 2021).

[20] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques third edition," *Morgan Kaufmann Ser. Data Manag. Syst.*, vol. 5, no. 4, pp. 83–124, 2011.

Received 22/June/2021; Accepted 26/September/2021