ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Prediction of Lung Cancer Disease Using Machine Learning Techniques

Rukhsar Hatam Qadir[1], Karwan Mohammed HamaKarim[2]

[1]Department of Statistics and Informatics, College-of-administration-and-economics, University of Sulaimani, Sulaimani, Kurdistan Region-Iraq, [2]Department of Information Technology, College of science and technology, University of Human Development, Sulaimani, Kurdistan Region-Iraq

ABSTRACT

The pursuit of algorithms utilizing external examples to formulate extensive hypotheses predicting the occurrence of novel instances is recognized, as supervised machine learning (SML). One of the jobs that intelligent systems perform the most frequently is supervised classification. The goal of this work is to evaluate supervised learning algorithms, explain SML classification methodologies, and identify the most effective classification algorithm given the available data. Two distinct machine learning (ML) techniques were examined: Random Forest (RF) and Neural Networks (NN). The algorithms were implemented using Python for knowledge analysis. For the categorization, 310 cases from a lung cancer data set were employed, with 15 features serving as independent variables and one serving as the dependent variable. In comparison to NN classification methods, RF was found to be the algorithm with the highest precision and accuracy, according to the results. The study reveals that while the kappa statistic and mean square error (MSE) are factors on the one hand, the time required to create a model and precision (accuracy) are factors on the other. Consequently, to have supervised predictive ML algorithms need to be precise, accurate, and minimum error. Thus, as a consequence of the research, we are currently at this analysis. The categorizing of NNs accuracy is 0.75 the MSE is 0.25, The RF classification accuracy is 0.89 and the MSE is 0.21.

Index Terms: Machine Learning, Classifiers, Data Mining Techniques, Data Analysis, Learning Algorithms, Supervised Machine Learning.

1. INTRODUCTION

Artificial intelligence (AI) is the capacity of a computer system or machine to perform tasks that normally require human intelligence. Description of two types of AI: Artificial narrow intelligence, or Weak AI, is the first and most prevalent type. Narrow intelligence, which encompasses all current AI systems, is task-specific and task-focused. The second is artificial general intelligence, which is the concept of a system having the capacity to think and act like a human (adaptable intellect) [1].

To put it simply, AI seeks to increase human capability and efficiency for activities, such as rebuilding nature and regulating society through intelligent machines, with the ultimate objective of achieving a society in which humans and machines live in harmony. Due to its long history, AI has been applied since the 1980s in several important fields, such as computer vision, natural language processing, the study of cognition and reasoning, robotics, game theory, and machine learning (ML) [2].

IT administration is entering a new era with the management of AI. To effectively manage AI, one must coordinate,

Access this article online

DOI:10.21928/uhdjst.v8n2y2024.pp75-83

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Rukhsar Hatam Qadir, Email: Rukhsar.qadir@univsul.edu.iq

Received: 02-04-2024 Accepted: 10-10-2024 Published: 17-11-2024

lead, communicate, and maintain control over a rapidly advancing field of computational innovations that draw on human ability to solve ever-more difficult decision-making challenges [3].

The past several years have witnessed a surge in interest in ML and AI due to the massive and continuous growth in data and processing capacity, as well as the development of better learning algorithms. Moreover, the notion that a computer might learn an abstract concept from data and use it in scenarios that haven't been seen before isn't new—it's been around at least since the 1950s. The community of clinical pharmacology and pharmacy cosmetics is well-versed in many of these fundamental concepts [4].

Healthcare applications of ML have drawn a lot of interest. Big data and increased processing capacity present a chance to employ ML algorithms to improve healthcare. The form of ML known as supervised learning may be used to forecast labeled data using support vector machines and methods, such as logistic or linear regression. Unsupervised ML models can recognize patterns in datasets that lack outcome information [5].

ML algorithms are especially useful for the healthcare sector because they enable us to make sense of the enormous volumes of medical data created daily in electronic health records. Finding patterns and insights in medical data that would be hard to detect manually can be aided by applying ML techniques, such as ML algorithms [6].

Data mining, also referred to as knowledge discovery in data, is a method employed to uncover patterns and valuable insights within extensive datasets. The adoption of data mining techniques, aiding organizations in transforming raw data into actionable information, has seen substantial growth in recent decades. This surge can be attributed in part to the expansion of big data and advancements in data warehousing technologies. Despite ongoing technological progress in managing large volumes of data, executives continue to encounter challenges related to automation and scalability [7].

Organizational decision-making has improved as a result of data mining's ability to generate insightful analyses of data. Data mining techniques underpin these investigations and have two main purposes: Either they characterize the target dataset or they apply data mining techniques [8].

Deep neural networks (NN), or multi-layered NNs, are the subject of deep learning, a branch of ML. These networks have demonstrated exceptional performance in complex tasks, often surpassing traditional algorithms in terms of accuracy and efficiency [9].

Despite their successes, NNs pose challenges, such as interpretability, overfitting, and the need for substantial amounts of labeled data for training. Researchers continue to explore techniques to address these issues, contributing to the ongoing evolution and refinement of NN architectures. As technology advances, NNs are expected to play an increasingly integral role in shaping the future of AI and ML applications [10].

ML includes supervised learning, in which an algorithm is taught on a labeled dataset. In this case, corresponding output labels accompany the input data. To enable the algorithm to produce predictions or classifications when faced with novel or unfamiliar data, it must first learn a mapping from input to output [11].

In supervised learning, the training dataset serves as a teacher to the algorithm. When the algorithm encounters input-output pairings, it modifies its internal parameters to minimize the difference between the expected and actual results. Well-known supervised ML methods include NNs, support vector machines (SVM) for classification difficulties, Random Forest (RF) for both regression and classification tasks, and Linear Regression for handling regression problems [12].

In this paper, we employ the RF algorithm and a NN, one kind of computer model that takes its cues from the structure and functions of the human brain is the NN. It is a key element of AI and ML, with the ability to identify patterns in data. NNs are fundamentally made up of layers of interconnected nodes, or artificial neurons. Typically, these layers are divided into three categories: Input, hidden, and output layers. Based on the input data and the intended output, the network learns by varying the weights of connections between neurons. This procedure, known as training, involves iterative optimization using algorithms, such as backpropagation.

NNs exhibit remarkable performance in tasks, such as speech and picture recognition, natural language processing, and even gaming. The capacity to extrapolate patterns from training data and offer predictions or classifications for previously unobserved data is their salient characteristic. NNs have applications in a wide range of sectors, from medical diagnostics to self-driving cars, because of their flexibility and adaptability [10].

Conversely, a RF is a well-liked ensemble learning method that is frequently used in ML for applications involving both regression and classification. The decision tree (DT) model is expanded by creating many DTs during training, and predictions are produced by averaging (for regression) or classifying (by majority vote) the individual trees [13].

The potential of ML to forecast lung cancer and change early detection and treatment approaches makes this research important. This work has the potential to greatly increase diagnostic accuracy by applying sophisticated algorithms to medical data, which could result in early interventions and better patient outcomes.

Furthermore, by addressing one of the primary causes of cancer-related mortality globally, this research has the potential to have a significant influence on public health and reduce healthcare expenses as well as the burden associated with late-stage diagnoses.

In this paper, we used Lung cancer disease, Lung cancer is a type of cancer that starts when abnormal cells grow uncontrolled in the lungs. It is a serious health issue that can cause severe harm and death.

Shortness of respiration, chest pain, and an ongoing cough are indicators of lung cancer. Early medical attention is important for preventing major health consequences. The course of medication is determined by the individual's medical history along with the disease's stage.

Small cell carcinoma (SCLC) and non-SCLC (NSCLC) are both of the most prevalent forms of lung cancer. While SCLC is less popular yet typically grows swiftly, NSCLC is more widespread and grows slowly.

Lung cancer is an important global problem of death and is also a significant public health issue. Based on the GLOBOCAN 2020 forecasts of cancer incidence and mortality published by the International Organization for Researching on Cancer, lung cancer remains the most prevalent cause of death due to cancer, contributing to 1.8 million deaths (18%) in 2020.

Tobacco usage, which involves applying cigars, pipes, and cigarettes, is the primary cause of lung cancer, yet it can also infect non-smokers. Further risk factors include a history of chronic lung conditions, exposure to second-hand tobacco smoke, inherited cancer conditions, air pollution, and certain chemicals and mesothelioma in the workplace [14].

2. RELATED WORK

In the study of Prakash *et al.* [15], to determine the age range that was most affected by the virus, a comprehensive analysis of COVID-19 data was conducted for this study. A number of prediction models are built using ML techniques, and their individual performances are computed and assessed. When compared to other ML models, such as SVM, K-Nearest Neighbor (KNN)+NCA, DT, Gaussian Naïve Bayesian, Multilinear, Logistic, and XGBoost classifiers, the RF Regressor and RF classifier both performed better.

Sharma *et al.* [16], in this investigation, utilize the DT, RF, and KNN ML prediction algorithms to examine how students do academically in relation to the amount of time they devote to extracurricular activities. In addition, an evaluation is carried out between the prediction outcomes attained by these various approaches to identify the underlying reasons for the shortcomings found in each machine-learning methodology. With an accuracy of 85% and an F1 84, the DT outscored nearly every other algorithm on our dataset.

In the study of Soni and Varma [17], ML classification and ensemble methods will be used to forecast diabetes using a dataset, which are KNN, DT, RF, GB, LR, and SVM. Each model is uniquely accurate in relation to other models. The project's output is a model which is as accurate – or perhaps more accurate – that is to show how well the model can predict diabetes. The results of the study show that RF performed with greater accuracy than other machine-learning techniques.

In the study of Al-Batah *et al.* [18], the work uses the excellent potential of machine-learning approaches for early CC prediction. To be more precise, three widely used choosing features and ranking algorithms have been used to determine the most significant features that support the diagnosis process. Furthermore, utilizing primary data consisting of five hundred photos, training, and thorough evaluation of eighteen distinct classifiers belonging to six learning strategies have been carried out. Furthermore, this issue of unequal class distribution—which frequently occurs in medical datasets—is studied. The results indicated that the Random Forest and LWNB classifiers performed best overall when four distinct assessment metrics were taken into account. In addition, logistic classifiers and LWNB demonstrated to be the most effective solutions for the common problem of unequal class distribution in the medical industry.

Abunasser *et al.* [19]'s research in this work, the performance of the instructors was predicted using a dataset from the University

of California at Irvine Repository. The efficiency of an instructor in colleges and universities was assessed by predicting their success using a variety of machine and deep learning techniques. The extra trees regressor scored 98.78%, 98.78%, 98.78%, and 98.78%, respectively, making it the greatest machine-learning algorithm when it comes to accuracy, precision, recall, and F1-score. The recommended deep learning methodology, on the other hand, received scores of 98.89%, 98.91%), 98.94%, and 98.92%.

3. CLASSIFICATION

Classification entails sorting a dataset into specific groups, a task relevant to both structured and unstructured data. The first step in this procedure is predicting the class of given data points. Terms, such as target labels and categories are used interchangeably for these classes. Predictive modeling describes the process of estimating the mapping function from discrete input variables to a discrete output variable. The main goal is to identify the category or class to which new data belongs [20].

The classification algorithm is a supervised learning technique that is used to identify the category of new observations based on training data. In classification, a program learns from the given dataset or observations and then classifies new observations into many classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, and cat or dog. Classes can be called targets/labels or categories.

The main goal of the classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.

In the below Fig. 1, there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes.

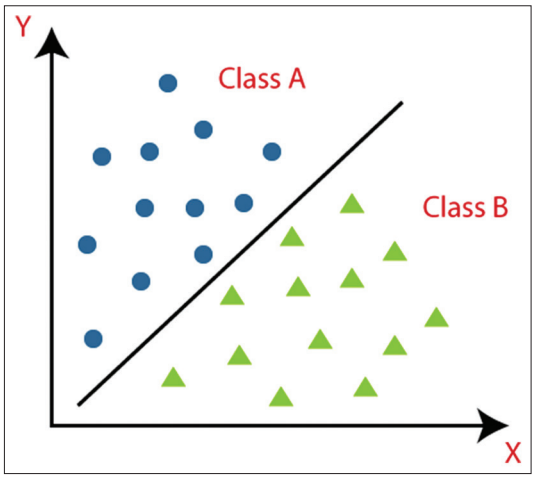

Fig. 1. Classification [21].

In this study, we utilized ML algorithms for lung cancer classification.

3.1. Measure of Classification

Numerous performance evaluation metrics exist for choosing a classification model, and using them effectively can lead to the creation of an optimal classification model.

The performance evaluation measures for classification models are:

- Confusion Matrix
- Precision
- Recall/Sensitivity
- Specificity
- F1-Score
- Area under curve (AUC) and receiver operating characteristics (ROC) curve.

3.1.1. Confusion matrix

In the context of a binary classifier, the confusion matrix [22] is illustrated in Fig. 2. The actual values can be either positive or negative, and the predictions categorize them as either positive or negative. The assessment of classification model probabilities is based on the terms true positive (TP), true negative (TN), false positive (FP), and false negative (FN) which exist in the confusion matrix. And Accuracy, Precision, and Recall are defined by:

TP, denoted as TP occurs in the confusion matrix when there is an expectation of a positive outcome, and the actual result aligns with the prediction.

FP In the confusion matrix, data points emerge when there is an anticipation of a positive outcome, but a negative event transpires. This situation constitutes a type 1 Error and can be likened to an unfortunate stroke of bad luck.

TP data points are found in the confusion matrix when a positive outcome is anticipated and the actual result matches the prediction.

FP, data points appear in the confusion matrix when a positive outcome is expected but a negative event occurs. A Type 1 Error is what happens in this situation. It is akin to a blessing in bad luck.

3.1.2. Accuracy

This phrase indicates the proportion of correct classifications among all categories. Stated otherwise, the number of TPs

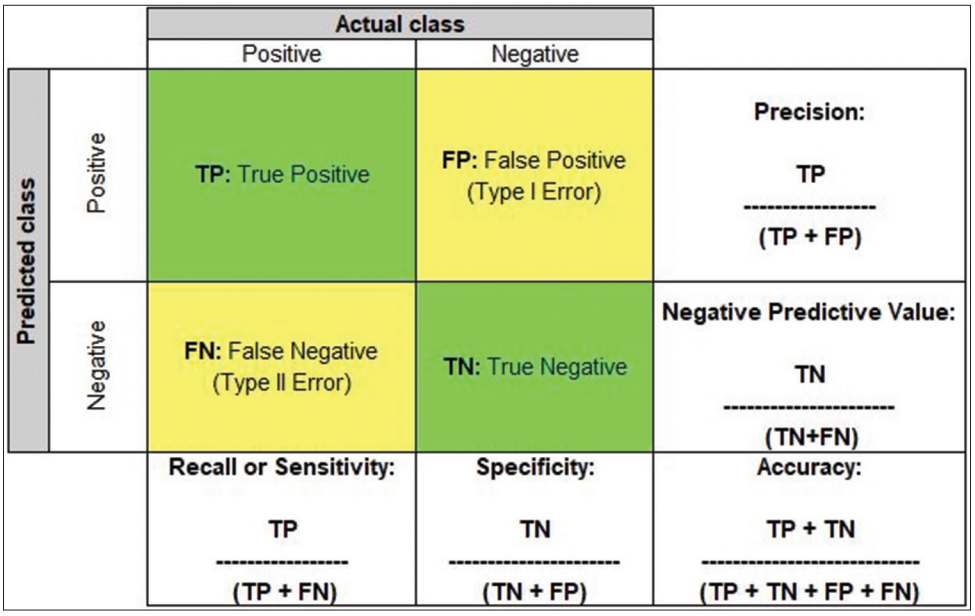

Fig. 2. Confusion matrix for the binary classification problem [23].

and TNs completed out of TP + TN + FP + FN. The ratio of "True" to the total of "True" and "False" is indicated.

Accuracy = (TP + TN)/(TP + FP +TN + FN)

3.1.3. Precision

What is the count of those identified as positive that genuinely belong to the positive category?

Precision = TP/(TP + FP)

3.1.4. Recall or sensitivity

Out of all the actual real positive cases, how many were identified as positive.

Recall = TP/(TN + FN)

3.1.5. Specificity

Out of all the real negative cases, how many were identified as negative.

Specificity = TN/(TN + FP)

3.1.6. F1-Score

F1 score = 2* (Precision * Recall)/(Precision + Recall)

3.1.7. AUC and ROC curve

AUC, which stands for the area under curve, is employed in conjunction with the ROC curve, also known as the ROC Curve. AUC represents the area under the ROC curve.

4. METHODOLOGY

The question definition recommends employing ML algorithms to forecast lung cancer, which addresses the urgent need for early detection and better treatment results. The aim of this project is to create predictive algorithms that can recognize patterns suggestive of the development of lung cancer by leveraging current information sources such as imaging scans and medical records. The goal of the research is to improve lung cancer screening efficiency and accuracy by concentrating on ML approaches. This may result in improved patient prognosis and early intervention. Within the study's purview is an evaluation of the viability of incorporating ML for routine screening and diagnosis into clinical practice. The ultimate goal of the research is to use AI to address a major public health issue, which would improve healthcare.

Using an ordinary dataset, two techniques for ML are implemented and compared in this work. The implementation's outcomes indicate each algorithm's precision in forecasting. The paragraph that follows describes the algorithms used in the present research.

- NN
- RF.

4.1. NN

The phrase "neural network" originated in the context of AI research, which sought to comprehend and imitate the

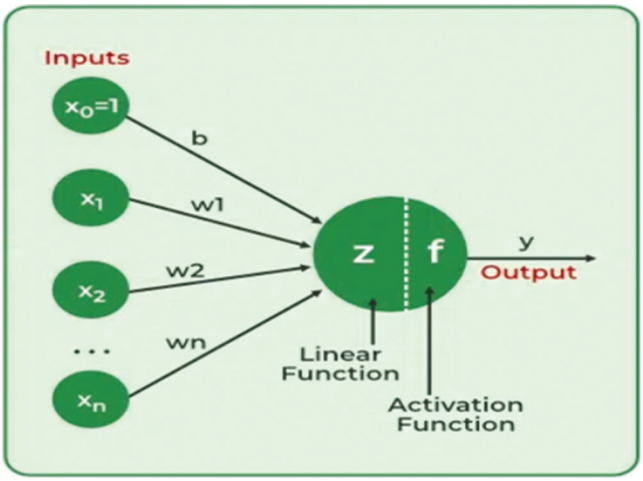

Fig. 3. Overview of a neural network's learning process [25].

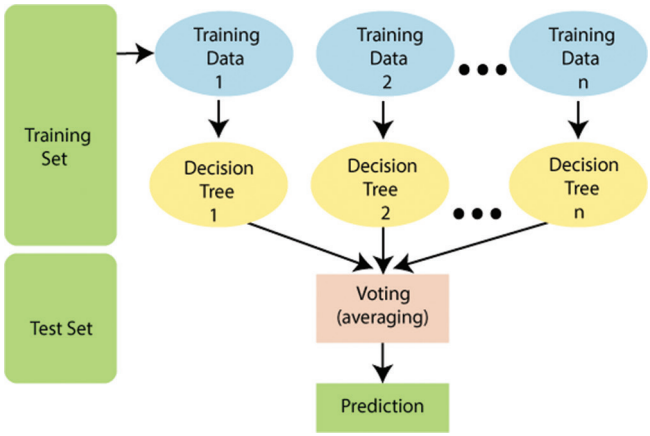

Fig. 4. Explains the working of the Random Forest algorithm [27].

workings of the human brain. AI is a field of computational science that focuses on creating intelligent machines, or computers with human-like thought and behavior. Artificial NNs have become a potent tool for solving hard problems in a variety of industries in recent years. An input layer, one or more hidden layers, and an output layer comprise an artificial NN. The hidden layers process inputs to generate the desired output. The learning algorithm that is employed impacts how well NNs succeed in tasks involving understanding patterns [24].

NNs have progressed to the point the fact that formerly unattainable jobs may now be completed with ease. Deeper association development in data sets, speech recognition, and image identification has all become more simpler.

A new context is used to mimic a NN's behavior. The simulation modifies the free parameters of the NN. As demonstrated by Fig. 3, the NN then adapts to the surroundings in an alternative manner as a result of the adjustments to its free parameters.

4.2. RF

In comparison to DTs, the RF is a more effective classifier. High-risk groups can be evaluated as a result of its numerous elements, which allow for early diagnosis at a stage that is treatable and curable. If somebody has smoked heavily in the past (more than 30 pack-years), is now a smoker, or has quit within the past 15 years, they have been classified as high-risk. Raising the number of cancer of the lungs screenings could save 30,000–60,000 lives annually in the United States, as the American Cancer Society predicts that 135,720 humans will be passing away from lung cancer in 2020. The U.S. Preventive Services Task Force has recommended lowering the age at which screening should begin (from 55 to 50 years old) and lowering the criteria for a smoking history (from 30 to 20 years old) [26].

The Fig. 4 explains the working of the RF algorithm.

5. DATASET DETAILS

The efficiency of a cancer prediction system enables individuals to assess their cancer risk affordably and facilitates informed decision-making based on their cancer risk status. The data is gathered from the online lung cancer prediction system's website.

Total no. of attributes: 16 No of instances: 310

Attribute information: (1) Gender: M (male), F (female) (2) Age: Age of the patient (3) Smoking: YES = 2, NO = 1. (4) Yellow fingers: YES = 2, NO = 1. (5) Anxiety: YES = 2, NO = 1. (6) Peer pressure: YES = 2, NO = 1. (7) Chronic disease: YES = 2, NO = 1. (8) Fatigue: YES = 2, NO = 1. (9) Allergy: YES = 2, NO = 1. (10). Wheezing: YES = 2, NO = 1. (11). Alcohol: YES = 2, NO = 1. (12) Coughing: YES = 2, NO = 1. (13) Shortness of breath: YES = 2, NO = 1. (14) Swallowing difficulty: YES = 2, NO = 1. (15) Chest pain: YES = 2, NO = 1. (16) Lung cancer: YES, NO.

5.1. Data Pre-Processing

Data preprocessing converts data into a format that may be used in ML, data mining, and other data science processes more quickly and effectively. To guarantee the generation of trustworthy results, these strategies are frequently applied early in the ML and AI development process.

Thus, data preparation transforms an unclean data set form initial information. Before the set of information is sent to the algorithm, it is preprocessed to look for values that are absent, noisy data, and numerous other inconsistencies. All data must be in ML-appropriate formats.

6. RESULTS AND DISCUSSION

This research makes a substantial contribution to the field by improving our understanding of how ML approaches, notably RF and NNs, might be used to detect lung cancer. The findings have significant practical consequences, as they point to improved early detection, which can lead to better patient outcomes and more effective therapies. Technologically, the work advances the field by demonstrating the RF algorithm's greater accuracy in predicting lung cancer. The broader public health benefit includes the potential for lower mortality rates and healthcare costs due to early diagnosis and intervention. Furthermore, the study sets the path for future studies that will integrate these models into clinical practice and investigate their applicability. The robustness and dependability of the results are tested.

6.1. NN Classification

The NN classification accuracy. The default values specified by the Python sci-kit-learn library package is used as initial values. As shown in Table 1 the accuracy, and mean square error (MSE) is 0.75, and 0.25, respectively.

The confusion matrix for the lung cancer disease dataset is shown in Fig. 5: each (TN, FN, FP, TP) are (28, 9, 10, 28), respectively.

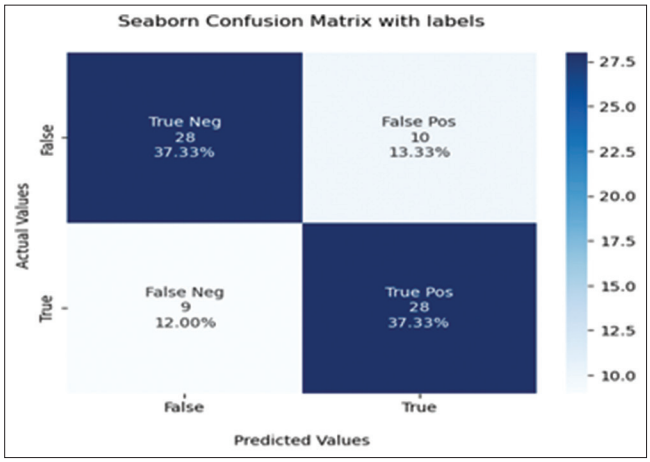

Fig. 5. Confusion matrix for the neural network.

Table 2 represents the comparison of multiclass classification reports (NN). The performance measure was calculated from the classification report.

The results show that the values for Precision, Recall, F1-score, and Support for Class 0 (which denotes lung cancer) are, respectively, 0.76, 0.74, 0.75, and 38. Similarly, the scores are 0.74, 0.76, 0.75, and 37 for Class 1 (showing no lung cancer). The accuracy as a whole is 0.75. Furthermore, the precision, recall, F1-score, and support values for the macro average are 0.75, 0.75, 0.75, and 75, respectively. Finally, the precision, recall, F1-score, and support weighted averages are 0.75, 0.75, 0.75, and 75.

6.2. RF Classification

The RF classification accuracy. The default values specified by the Python sci-kit-learn library package is used as initial values. As shown in Table 3 the accuracy, and MSE is 0.89, and 0.21, respectively.

The confusion matrix for the lung cancer disease dataset is shown in Fig. 6: each (TN, FN, FP, TP) are (1, 0, 8, 66), respectively.

Table 4 represents the comparison of multiclass classification reports (Random Forest). The performance measure was calculated from the classification report.

The value of accuracy is 0.89; additionally, the values of Precision, recall, f1-score, and Support are 0.95, 0.56, 0.57, and 75, respectively, for macro average and 0.90, 0.89, 0.85, and 75, respectively, for weight average. The results show that the values of Precision, recall, f1-score, and Support are 1.00, 0.11, 0.20, and 9, respectively, for Class 0 (Has lung Cancer) and 0.89, 1.00, 0.94, 66, respectively, for Class 1 (Has not lung Cancer).

TABLE 1: Performance comparison of neural network

Dataset	Accuracy	Mean square error
Lung cancer	0.75	0.25

TABLE 2: Represents the comparison of a multiclass classification report (neural network)

	Precision	Recall	F1-score	Support
0	0.76	0.74	0.75	38
1	0.74	0.76	0.75	37
accuracy			0.75	75
Macro avg	0.75	0.75	0.75	75
Weighted avg	0.75	0.75	0.75	75

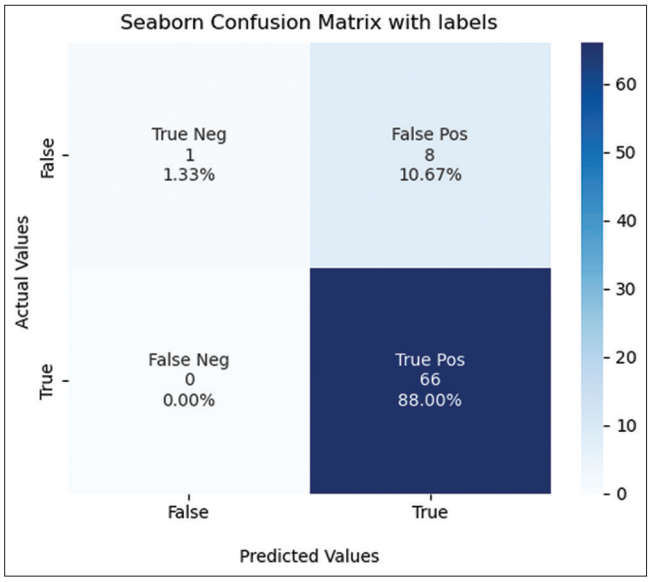

Fig. 6. Confusion matrix for Random Forest.

TABLE 3: Performance comparison of Random Forest

Dataset	Accuracy	Mean square error
Lung cancer	0.89	0.21

TABLE 4: Represents the comparison of a multiclass classification report (Random Forest)

	Precision	Recall	F1-score	Support
0	1.00	0.11	0.20	9
1	0.89	1.00	0.94	66
accuracy			0.89	75
Macro avg	0.95	0.56	0.57	75
Weighted avg	0.90	0.89	0.85	75

7. CONCLUSION

Fine-tuning parameters and having a substantial number of instances in the dataset are essential aspects of ML classification. Constructing the model for an algorithm involves not only time but also focuses on precision and accurate classification. It's important to note that the effectiveness of a learning algorithm for one dataset does not guarantee precision and accuracy for another dataset with logically different attributes.

The central question in ML classification isn't about the superiority of one learning algorithm over others but about identifying the conditions under which a specific method can outperform others for a given application problem. Meta-learning addresses this concern by attempting to find functions that map datasets to algorithm performance.

Once the strengths and weaknesses of each approach are comprehended, it becomes essential to investigate the potential of combining two or more algorithms to address a problem. The objective is to capitalize on the strengths of one method to offset the shortcomings of another. Striving for optimal classification accuracy can be challenging when seeking a single classifier that matches the performance of a proficient ensemble of classifiers. ML algorithms such as NNs and RFs can provide elevated precision and accuracy, irrespective of the number of attributes and data instances.

This research emphasizes that the time required to build a model is one factor, and precision with metrics, such as kappa statistics and MSE is another. Therefore, ML algorithms demand precision, accuracy, and minimal error for effective supervised predictive ML. For large datasets, a recommendation is made to consider a distributed processing environment, as it allows for a high correlation among variables, ultimately enhancing the efficiency of the model output.

The results have shown that for lung disease prediction, the RF algorithm performs greater than the NN when it comes of precision and reliability. This implies that the application of ML techniques has an enormous opportunity to enhance early diagnosis and treatment plans. It should be the focus of future research to demonstrate the practical value of these models in healthcare environments and improve patient care.

Therefore, as a result of the study, we have come to this analysis‘ The NN classification accuracy is 0.75 the MSE is 0.25, The RF classification accuracy is 0.89 and the MSE is 0.21.

REFERENCES

[1] C. L. Hann, M. A. Wu, N. Rekhtman and C. M. Rudin. In: V. T. DeVita, T. S. Lawrence and S. A. Rosenberg. "*Cancer Principles and Practice of Oncology*". Ch. 49. Wolters Kluwer, Netherlands, pp. 671-700, 2019.

[2] J. Liu, X. Kong, F. Xia, X. Bai, L. Wang, Q. Qing and I. Lee. "Artificial intelligence in the 21st century". *IEEE Access*, vol. 6, pp.34403-34421, 2018.

[3] B. Mahesh. "Machine learning algorithms-a review". *International Journal of Science and Research*, vol. 9, no. 1, pp. 381-386, 2020.

[4] S. Badillo, B. Banfai, F. Birzele, I. I. Davydov, L. Hutchinson, T. Kam-Thong, and Zhang. An introduction to machine learning. Clinical pharmacology & therapeutics, vol. 107. no. 4, pp. 871-885, 2020.

[5] A. Alanazi. "Using machine learning for healthcare challenges and opportunities". *Informatics in Medicine Unlocked*, vol. 30, p. 100924, 2022.

[6] Available from: https://www.foreseemed.com/blog/machine-learning-in-healthcare [Last accessed on 2024 Jul 12].

[7] J. Wang, Q. Liu, S. Yuan, W. Xie, Y. Liu, Y. Xiang, N. Wu, L. Wu, X. Ma, T. Cai, Y. Zhang, Z. Sun and Y. Li. "Genetic predisposition to lung cancer: Comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies". *Scientific Reports*, vol. 7. p. 8371, 2017.

[8] K. Ten Haaf, C. M. Van der Aalst, H. J. De Koning, R. Kaaks and M. C. Tammemägi. "Personalising lung cancer screening: An overview of risk-stratification opportunities and challenges". *International Journal of Cancer*, vol. 149, no. 2, pp. 250-263, 2021.

[9] I. Toumazis, M. Bastani, S. S. Han and S. K. Plevritis. "Risk-based lung cancer screening: A systematic review". *Lung Cancer*, vol. 147, pp. 154-186.

[10] A. S. Tsao, G. V. Scagliotti, P. A. Bunn Jr., D. P. Carbone, G. W. Warren, C. Bai, H. J. De Koning, A. U. Yousaf-Khan, A. McWilliams, M. S. Tsao, P. S. Adusumilli, R. Rami-Porta, H. Asamura, P. E. Van Schil,… & H. I. Pass. "Scientific advances in lung cancer 2015". *Journal of Thoracic Oncology*, vol. 11, no. 5, pp. 613-638, 2016.

[11] K. Ten Haaf, J. Jeon, M. C. Tammemägi, S. S. Han, C. Y. Kong, S. K. Plevritis, E. J. Feuer, H. J. De Koning, E. W. Steyerberg and R. Meza. "Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study". *PLoS Medicine*, vol. 14, no. 4, p. e1002277, 2017.

[12] J. Sands, M. C. Tammemägi, S. Couraud, D. R. Baldwin, A. Borondy-Kitts, D. Yankelevitz, J. Lewis, F. Grannis, H. U. Kauczor, O. Von Stackelberg, L. Sequist, U. Pastorino and B. McKee. "Lung screening benefits and challenges: A review of the data and outline for implementation". *Journal of Thoracic Oncology*, vol. 16, no. 1, pp.37-53, 2021.

[13] H. A. Katki, S. A. Kovalchik, L. C. Petito, L. C. Cheung, E. Jacobs, A. Jemal, C. D. Berg and A. K. Chaturvedi. "Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening". *Annals of Internal Medicine*, vol. 169, no. 1, pp. 10-19, 2018.

[14] Available from: https://www.who.int/news-room/fact-sheets/detail/lungcancer?gad_source=1&gclid=cj0kcqjwhb60bhclarisabggtw9cqlaualhejgxrilmmi478y3jo5hxkyyvg0zkwxctc9u3kozdubagaatmeealw_wcb [Last accessed on 2024 Jul 10].

[15] K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar and Y. N. Pawan. "Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms". *International Journal*, vol. 8, no. 5, pp. 2199-2204, 2020.

[16] N. Sharma, S. Appukutti, U. Garg, J. Mukherjee and S. Mishra. "Analysis of student's academic performance based on their time spent on extra-curricular activities using machine learning techniques". *International Journal of Modern Education and Computer Science*, vol. 15, no. 1, pp. 46-57, 2023.

[17] M. Soni and S. Varma. "Diabetes prediction using machine learning techniques". *International Journal of Engineering Research and Technology*. Vol. 9, no. 9, pp. 2278-0181, 2020.

[18] M. S. Al-Batah, M. Alzyoud, R. Alazaidah, M. Toubat, H. Alzoubi and A. Olaiyat. "Early prediction of cervical cancer using machine learning techniques". *Jordanian Journal of Computers and Information Technology*, vol. 8, no. 4, p. 1, 2022.

[19] B. S. Abunasser, M. R. J. AL-Hiealy, A. M. Barhoom, A. R. Almasri and S. S. Abu-Naser. "Prediction of instructor performance using machine and deep learning techniques". *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022.

[20] N. Tariq. "Breast cancer detection using artificial neural networks". *Journal of Molecular Biomarkers and Diagnosis*, vo. 9, no. 1, pp. 371, 2017.

[21] Available from: https://www.javatpoint.com/classification-algorithm-in-machine-learning [Last accessed on 2024 Jul 10].

[22] Ž. Vujović. "Classification model evaluation metrics". *International Journal of Advanced Computer Science and Applications*, vol. 12, no, 6, pp. 599-606, 2021.

[23] Available from: https://www.analyticsvidhya.com/blog/2020/12/decluttering-the-performance-measures-of-classification-models [Last accessed on 2024 Jul 10].

[24] R. Dharwal and L. Kaur. "Applications of artificial neural networks: A review". *Indian Journal of Science and Technology*, vol. 9, no. 47, p. 1-8, 2016.

[25] Available from: https://www.geeksforgeeks.org/neural-networks-a-beginners-guide [Last accessed on 2024 Jul 10].

[26] M. Ismail. "Lung cancer prediction using data mining techniques". *International Journal of Recent Technology and Engineering*, vol. 8, no. 4, pp. 12301-12305, 2019.

[27] Available from: https://www.javatpoint.com/machine-learning-random-forest-algorithm [Last accessed on 2024 Jul 10].