engineering proceedings

MDPI

Proceeding Paper

Early Detection of Mesothelioma Using Machine Learning Algorithms [†]

Taimur Shahzad Gill *, **Muhammad Ayaz Shirazi** * and **Syed Sajjad Haider Zaidi**

Department of Electronics and Power Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan; sajjadzaidi@pnec.nust.edu.pk

* Correspondence: taimur.beee19pnec@student.nust.edu.pk (T.S.G.); ayaz.shirazi@pnec.nust.edu.pk (M.A.S.)

[†] Presented at the 8th International Electrical Engineering Conference, Karachi, Pakistan, 25–26 August 2023.

Abstract: Early detection of mesothelioma, a severe form of cancer commonly associated with asbestos exposure, is a significant challenge that greatly affects prognosis. This study addresses this issue using Machine Learning (ML) algorithms, including Gradient-Boosted Trees (GBT), Support Vector Machines (SVM), and Logistic Regression (LR). This study uses the mesothelioma dataset from the UCI Machine Learning Repository to evaluate the proposed models, achieving 100% accuracy and F1 score in detecting the disease, accurately classifying 98 patients with 30 true positives and 68 true negatives. Further analysis using the AUC-ROC score showed that the 'duration of symptoms' feature was most informative for the GBT model, with a score of 0.595. In contrast, 'C-reactive protein' was the most significant feature for the SVM and LR models, each achieving an AUC-ROC of 0.603. Despite these promising results, validating these findings with additional datasets is critical to confirm their generalisability. However, this study provides strong support for using ML algorithms in the early detection of mesothelioma, potentially leading to improved patient diagnoses.

Keywords: mesothelioma; machine learning; diagnosis; gradient-boosted trees; support vector machines; logistic regression; receiver operating characteristics

check for updates

Citation: Gill, T.S.; Shirazi, M.A.; Zaidi, S.S.H. Early Detection of Mesothelioma Using Machine Learning Algorithms. *Eng. Proc.* **2023**, *46*, 6. https://doi.org/10.3390/engproc2023046006

Academic Editors: Abdul Ghani Abro and Saad Ahmed Qazi

Published: 20 September 2023

1. Introduction

Mesothelioma, a malignant tumour primarily associated with asbestos exposure, affects the thin tissue layer covering most of our internal organs—the mesothelium [1]. With nearly 3000 new cases every year in the United States alone, the disease's impact is significant, mainly due to its aggressive nature and poor prognosis [2]. A challenge to early detection is the disease's long latency period, which can extend up to 50 years. Hence, there is a pressing need for innovative approaches to improve the diagnostic process [3]. This study explores the application of Machine Learning (ML) in addressing this challenge using three algorithms: Gradient-Boosted Trees (GBT), Support Vector Machines (SVM), and Logistic Regression (LR). While these algorithms have shown robust performance in various healthcare applications, their use in mesothelioma detection remains relatively unexplored [4].

2. Literature Review

Historically, the detection of mesothelioma has relied heavily on imaging techniques and invasive procedures like biopsies. According to Pass et al. [5], high-resolution CT scanning and MRI have greatly improved the accuracy of diagnosing mesothelioma. However, these techniques often only identify the disease in its later stages, limiting the potential for successful treatment. Blood biomarkers have emerged as a promising non-invasive technique for early detection of mesothelioma. A study by Creaney and Robinson [6] showed the potential of Soluble Mesothelin-Related Peptide in detecting mesothelioma. Despite this progress, there is still a need for more precise and accessible diagnostic tools.

Machine learning has proven to be a powerful tool in healthcare. In a comprehensive review, Obermeyer and Emanuel [7] noted the significant potential of ML to augment every aspect of healthcare, from patient diagnosis to personald care. The application of ML algorithms in cancer detection has been widely studied. For example, a study by Cruz and Wishart [8] demonstrated the successful use of SVM in breast cancer detection. Meanwhile, Caruana et al. [9] successfully used GBT to predict pneumonia risk. However, the use of these algorithms in mesothelioma detection is still under-researched. Recent research has shed light on innovative diagnostic techniques for mesothelioma, emphasizing non-invasive methods. Brusselmans et al. [10] proposed breath analysis as a potential tool for detecting malignant pleural mesothelioma. This breath-based diagnostic method uses a specific profile of volatile organic compounds (VOCs) exhaled by patients. However, while promising, breath analysis is still an emerging technique and needs more validation for routine clinical application. Bononi et al. [11] demonstrated the potential of convolutional neural networks (CNNs) for classifying histopathological images of mesothelioma. This study highlights the promise of ML, particularly deep learning (DL), in aiding pathological diagnoses. Despite these advances, the application of ML algorithms specifically for mesothelioma detection is still limited, stressing the importance of present research.

3. Materials and Methods

This study uses KNIME to explore the potential application of ML algorithms for early mesothelioma detection [12]. A systematic approach was used based on the three ML models: GBT, SVM, and LR. Hyperparameters for each model were fine-tuned to optim performance, and their usefulness was then measured through a combination of evaluation metrics, ensuring both reliability and precision.

3.1. Mesothelioma Dataset

This study uses patient hospital reports from Dicle University's Faculty of Medicine to perform the research reported [9]. The dataset comprises 324 mesothelioma patient data, which have been diagnosed and treated, subsequently investigated retrospectively and analysed. Each sample in the dataset has 34 features. This feature selection was guided by medical expertise, asserting that this set of features is more effective in capturing the complexity of the disease. Some of the clinical features included are the type of mesothelioma, duration of asbestos exposure, duration of symptoms, chest pain, weakness, white blood cell count (WBC), haemoglobin (HGB), platelet count (PLT), total protein level, C-reactive protein (CRP), and diagnosis class. The diagnostic test results of each patient were carefully recorded, providing an extensive and in-depth set of data for analysis.

3.2. Data Preprocessing

A workflow was developed using the KNIME Analytics Platform (shown in Figure 1), ensuring that the data were prepared for the subsequent stages of analysis. One of the primary steps in preprocessing involved handling missing data within the records. Records containing missing values were excluded from the dataset to maintain data purity and integrity. Although this approach could potentially lead to a reduction in the sample size, it helped to prevent the introduction of potential bias or inaccuracies that could compromise the performance of the ML models.

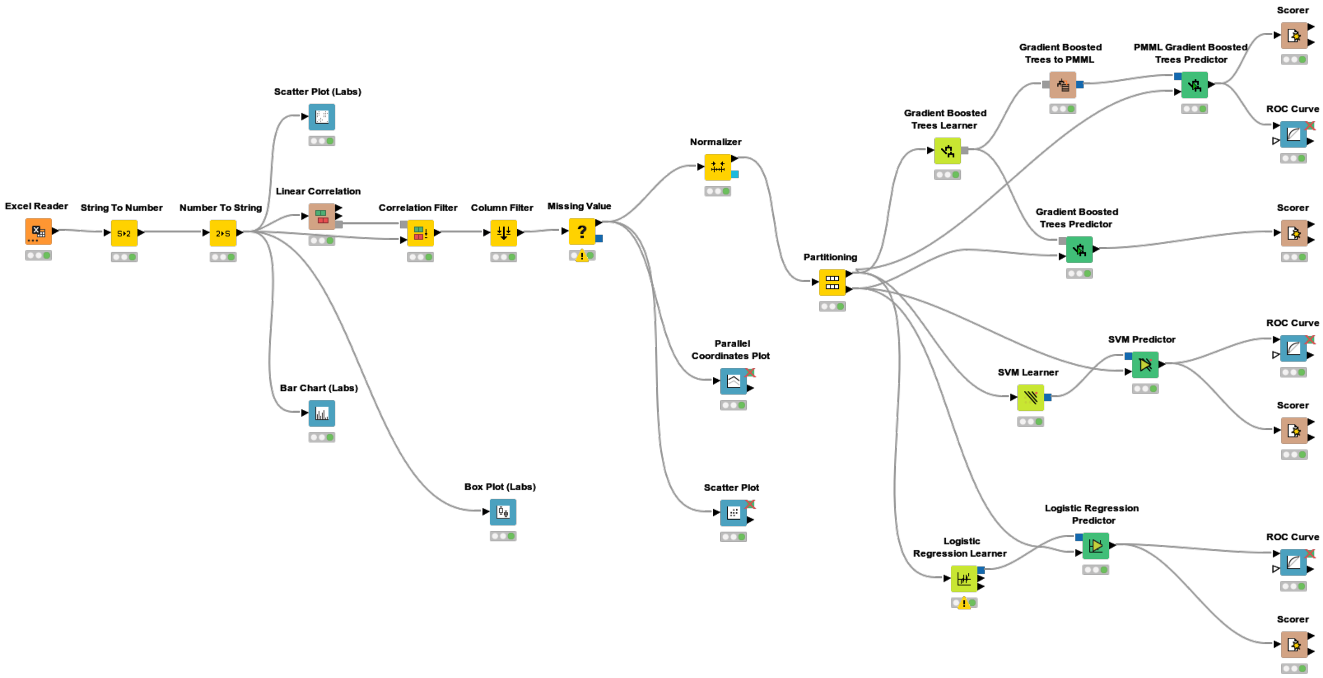

Figure 1. KNIME workflow for data preprocessing, visualisation, and mining.

3.3. Feature Selection

The raw dataset consisted of 34 features, each providing unique information about the patient's medical condition. However, having many features can sometimes lead to complications, such as multicollinearity and overfitting. Pearson's Correlation Coefficient (PCC) was used to examine the correlation between the features. Three features with PCC higher than 0.65 were identified and subsequently excluded from the dataset to avoid the risk of overfitting [13,14]. As a result, 31 features were selected for the model training, simplifying the dataset while retaining its capacity to yield reliable and valid results.

3.4. Data Mining

Data mining makes predicting future trends and behaviours possible, allowing businesses to make proactive, knowledge-driven decisions using sophisticated algorithms [15]. This study uses data mining to uncover hidden patterns and relationships within the mesothelioma dataset. The chosen algorithms— GBT, LR, and SVM—are renowned for their efficiency in handling large datasets and their ability to model partially complex relationships.

3.4.1. Gradient Boosted Trees

GBT is a powerful ML algorithm that builds an ensemble of decision trees sequentially, each correcting the errors made by the previous ones [16]. The learning process involves a loss function, a weak learner, and an additive model. The loss function measures the discrepancy between the predicted and actual outcomes. For GBT, the loss function is differentiable, and the weak learner is a decision tree. The model's hyperparameters were optimised for best performance; the tree depth was set to 4, the number of models was 100, and the learning rate was set to 0.1. The tree depth is the maximum length from the tree root to a leaf, and it determines the complexity of the model. A smaller number of trees can result in an underfit model, while a large number can lead to overfitting. The learning rate shrinks the contribution of each tree to prevent overfitting. Combining a lower learning rate with many trees is recommended by [17,18].

A logistic loss function was used with the XGBoost classifier, as shown in Equation (1):

$$y_{hat} = \sigma(F_M(x)) = \sigma(F_0(x) + v\sum_{m=1}^{M}(\sum_{j=1}^{J}\gamma_j m I(x \text{ in } R_{jm}) - \lambda\sum_{m}\Omega(h_m))) \tag{1}$$

Here, $\sigma(z) = 1/(1 + exp(-z))$ is the logistic function, $F_M(x)$ is the final boosted model, $F_0(x)$ is the initial model, v is the learning rate, M is the number of trees, J is the number of leaves in each tree, $\gamma_j m$ are the weights for each leaf, and $I(x \text{ in } R_{jm})$ is an indicator function that is 1 if the sample x is in the region (leaf) R_{jm} of the m-th tree and 0 otherwise. The additional term $-\lambda\sum_m \Omega(h_m)$ is a regularisation term where γ is the regularisation parameter, and $\Omega(h_m)$ is a complex function of the tree h_m. The complexity function can take a form like $\gamma T + 0.5\lambda||w||^2$, where T is the number of leaves in the tree, γ is the complexity control parameter for the number of leaves, w are the weights assigned to each leaf, and $||w||^2$ is the $L2$ norm of the weights. This additional term helps to control the complexity of the model and thus helps to reduce overfitting.

3.4.2. Logistic Regression

LR is a commonly used statistical model with a logistic function to model a binary dependent variable. It is a form of regression analysis where the dependent variable is categorical [19]. The logistic function, also called the sigmoid function, can take any real input t, as defined in Equation (2):

$$S(t) = 1/(1 + e^{-t}) \tag{2}$$

The output of $S(t)$ lies between 0 and 1, which is interpretable as a probability. In logistic regression, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. The logistic regression equation is shown in Equation (3):

$$P(Y = 1|X) = 1/(1 + e^{-\beta_0 + \beta_1 \times X}) \tag{3}$$

where $P(Y = 1|X)$ is the conditional probability that $Y = 1$ given the values of the explanatory variables X, β_0 is the intercept, and β_1 is the coefficient for the explanatory variable. The logistic regression model was trained, and the stochastic average gradient solver was selected for optimisation. This solver is effective for large datasets and computes an average of the gradient over several past updates, providing better generalisation and faster convergence.

3.4.3. Support Vector Machines

SVM is a supervised ML algorithm which seeks to find the best hyperplane that separates different classes in a high-dimensional space. SVM operates based on maximizing the margin around the separating hyperplane, contributing to better generalisation [20]. The SVM model visualises training examples as spatial points arranged so that a wide, distinct separation exists between different class examples for optimal classification. The core of the SVM algorithm is a quadratic programming problem seeking to find the separating hyperplane. The problem can be formulated as shown in Equation (4):

$$minimise\ (1/2) \times ||w||^2\ subject\ to\ y_i(w.xi - b) >= 1,\ i = 1, \ldots, n \tag{4}$$

where w is a weight vector, b is a bias term, and x_i and y_i are the training samples and their corresponding labels. A hyper-tangent kernel was used for SVM, with kappa set to 0.1, delta set to 0.5, and the overlapping penalty set to 1.

4. Results and Discussion

The experimental results revealed exceptional performance, as shown in Table 1. Each model achieved accuracy, precision, and recall values of 100% on the test dataset. All three models correctly classified all 98 instances in the dataset, with 30 true positives and

68 true negatives, resulting in no false positives or false negatives. Further analysis of the model's performance using the area under the receiver operating characteristic curve (AUC-ROC) showed a distinct picture of the models' performances. The GBT model achieved the highest AUC-ROC for the 'Duration of Symptoms' feature, scoring 0.595. On the other hand, both the SVM and LR models performed best with the 'C-reactive protein' feature, each achieving an AUC-ROC of 0.603. The exceptional performance of the three models is a promising sign of the potential of ML in mesothelioma detection. However, achieving perfect metrics raises concerns about potential overfitting and, thus, the need for further validation.

Table 1. Model comparison in terms of performance metrics.

Model	Accuracy	F1 Score	AUC ROC	MCC [1]
GBT	1.00	1.00	0.595	1.00
LR	1.00	1.00	0.603	1.00
SVM	1.00	1.00	0.603	1.00

[1] Mathew's correlation coefficient.

The high feature importance scores for 'Duration of Symptoms' and 'C-reactive protein', as shown in Figure 2, imply that these features are particularly informative for mesothelioma detection. Clinicians may want to pay particular attention to these symptoms during patient evaluations. This study's results provide significant evidence supporting the use of ML algorithms for the early detection of mesothelioma. It highlights the effectiveness of GBT, SVM, and LR models, highlighting the relevance of specific patient symptoms in prediction accuracy. This approach has the potential to significantly improve early detection rates and, consequently, patient outcomes.

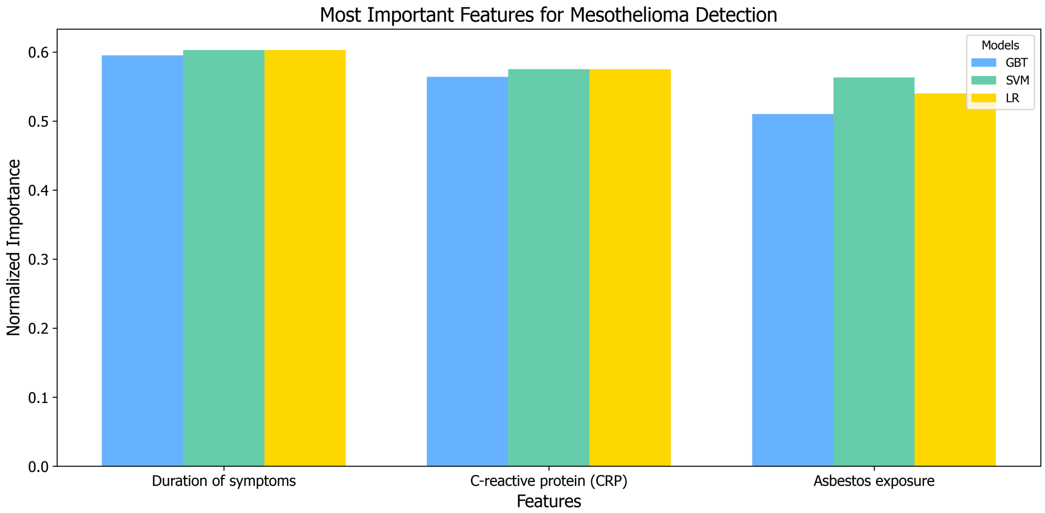

Figure 2. The three most important features for the proposed models (sorted from left to right in terms of decreasing predictive power).

5. Conclusions and Future Work

The application of ML algorithms in healthcare has demonstrated immense potential, as illustrated by the performance of the GBT, SVM, and LR models in this study. Each of these models achieved 100% accuracy, precision, and recall in detecting mesothelioma using the UCI mesothelioma dataset, pointing to the significant potential of these algorithms in this critical area of healthcare. While the results are promising, it is important to note that ML models should serve as tools to assist healthcare professionals rather than definitive diagnostic systems. They are not designed to replace traditional diagnostic methods but to augment them, providing an additional layer of information to support healthcare providers in making informed decisions. Despite the promising results, there are several avenues for further research. For one, validating these models with larger sample sizes and more diverse patient cohorts would be valuable to ensure generalisability. Additionally, testing the models with datasets featuring a larger variety of patients, more samples, and potentially noisier data would provide a more comprehensive evaluation of their effectiveness. Moreover, including other ML models could also enhance the robustness and accuracy of mesothelioma detection. DL algorithms, for example, which have shown significant success in image recognition tasks, could be applied to analyse medical imaging data such as CT scans or MRIs for early mesothelioma detection.

Author Contributions: Conceptualisation, T.S.G. and M.A.S.; methodology, T.S.G.; software, T.S.G.; validation, T.S.G., M.A.S. and S.S.H.Z.; formal analysis, T.S.G.; investigation, T.S.G.; resources, M.A.S. and S.S.H.Z.; data curation, T.S.G. and M.A.S.; writing—original draft preparation, T.S.G.; writing—review and editing, T.S.G., M.A.S. and S.S.H.Z.; visualisation, T.S.G.; supervision, M.A.S. and S.S.H.Z.; project administration, T.S.G. and S.S.H.Z.; funding acquisition, M.A.S. and S.S.H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analysed in this study. This data can be found here: https://archive.ics.uci.edu/ml/machine-learning-databases/00351/ (accessed on 21 May 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baumann, F.; Carbone, M. Environmental risk of mesothelioma in the United States: An emerging concern—Epidemiological issues. *J. Toxicol. Environ. Health Part B* **2016**, *19*, 231–249. [CrossRef] [PubMed]
2. Teta, M.J.; Mink, P.J.; Lau, E.; Sceurman, B.K.; Foster, E.D. US mesothelioma patterns 1973–2002: Indicators of change and insights into background rates. *Eur. J. Cancer Prev.* **2008**, *17*, 525–534. [CrossRef] [PubMed]
3. Bibby, A.C.; Tsim, S.; Kanellakis, N.; Ball, H.; Talbot, D.C.; Blyth, K.G.; Maskell, N.A.; Psallidas, I. Malignant pleural mesothelioma: An update on investigation, diagnosis and treatment. *Eur. Respir. Rev.* **2016**, *25*, 472–486. [CrossRef] [PubMed]
4. Deo, R.C. Machine learning in medicine. *Circulation* **2015**, *132*, 1920–1930. [CrossRef] [PubMed]
5. Pass, H.I.; Levin, S.M.; Harbut, M.R.; Melamed, J.; Chiriboga, L.; Donington, J.; Huflejt, M.; Carbone, M.; Chia, D.; Goodglick, L.; et al. Fibulin-3 as a blood and effusion biomarker for pleural mesothelioma. *N. Engl. J. Med.* **2012**, *367*, 1417–1427. [CrossRef]
6. Creaney, J.; Robinson, B.W. Serum and pleural fluid biomarkers for mesothelioma. *Curr. Opin. Pulm. Med.* **2009**, *15*, 366–370. [CrossRef]
7. Obermeyer, Z.; Emanuel, E.J. Predicting the future—Big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **2016**, *375*, 1216. [CrossRef] [PubMed]
8. Cruz, J.A.; Wishart, D.S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2006**, *2*, 117693510600200030. [CrossRef]
9. Caruana, R.; Karampatziakis, N.; Yessenalina, A. An empirical evaluation of supervised learning in high dimensions. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 96–103.
10. Brusselmans, L.; Arnouts, L.; Millevert, C.; Vandersnickt, J.; van Meerbeeck, J.P.; Lamote, K. Breath analysis as a diagnostic and screening tool for malignant pleural mesothelioma: A systematic review. *Transl. Lung Cancer Res.* **2018**, *7*, 520. [CrossRef] [PubMed]
11. Bononi, I.; Comar, M.; Puozzo, A.; Stendardo, M.; Boschetto, P.; Orecchia, S.; Libener, R.; Guaschino, R.; Pietrobon, S.; Ferracin, M.; et al. Circulating microRNAs found dysregulated in ex-exposed asbestos workers and pleural mesothelioma patients as potential new biomarkers. *Oncotarget* **2016**, *7*, 82700. [CrossRef] [PubMed]
12. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e. V., Freiburg, Germany, 7–9 March 2007.
13. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 2.
14. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 112.

15. Ramamohan, Y.; Vasantharao, K.; Chakravarti, C.K.; Ratnam, A. A study of data mining tools in knowledge discovery process. *Int. J. Soft Comput. Eng. (IJSCE) ISSN* **2012**, *2*, 2231–2307.
16. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
17. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
18. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobotics* **2013**, *7*, 21. [CrossRef] [PubMed]
19. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013; Volume 398.
20. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.