**REVIEWER COMMENTS ADDRESSING FORM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Corresponding Author Name** | | Luxshi Karunakaran | | |
| **EDAS ID** | | 1571137894 | | |
| **Research Title** | | Enhancing Lung Cancer Prediction Using Machine Learning: A Comparative Analysis of Hyperparameter Optimization Techniques | | |
| **Authors/co-authors** | | Luxshi Karunakaran  Senthan Prasanth  Chandrika Malkanthi  Rathnayaka Mudiyanselage Kapila Tharanga Rathnayaka | | |
|  | | | | |
| **Review No.** | **Review Comment** | | **Addressed Solution** | **Section No. Page No. Line No.** |
| 1 | The specific novel contribution of this work needs to be more explicitly highlighted. Many studies exist on ML/DL for lung cancer prediction and hyperparameter tuning. What makes this comparative analysis stand out? | | The study will be unique among other ML and DL studies on lung cancer prediction due to its ability to introduce an extensive system combining Bayesian optimization with hyperparameter tuning of ML and DL models for numerical data. In contrast, most of the existing studies main focus is lung cancer pattern identification using images with straightforward validation schemes. This pa-  per will discuss the effect of different cross-validation methods (hold-out, k-fold, stratified k-fold, LOO) on numerical data prediction with a strong emphasis on increasing both accuracy and efficiency of the model. The model’s ability to accurately diagnose the cancer using numeric data aid to increases the accessibility of health diagnostics in the underprivileged regions, ensuring healthcare equity. Moreover, the paper addresses another limitation in the literature by providing a comprehensive performance assessment system with a positive trade-off between performance and computational demands, focusing accuracy, precision, recall, F1-score, AUC-ROC, confusion matrix, and training time. This has been substantiated by the fact that the prediction has significantly improved where its hyperparameters have been Bayesian tuned; hence, it is now a scalable and efficient model which is ready to be integrated into clinical practice, therefore representing a reference point to predict lung cancer through numerical data. | Section [2.1],  Page [5],  Line [2] |
| Provide a detailed description of the 16 numerical attributes used, including their types, ranges, and any specific characteristics relevant to lung cancer prediction. | | Added the Dataset description as a table. | Section No [3.1]  Page no [7]  Line no [1] |
| Clearly describe the methods used for data preprocessing. | | In addition to mean imputation of missing values, the two important preprocessing methods are encoding categorical variables and feature scaling. Encoding transforms the categories, such as gender (M/F) and lung cancer (YES/NO) into numbers (label encoding) so that they can be compatible with algorithms. Numerical features such as age are normalized by feature scaling to balance model training. These procedures enhance the quality of data. | Section No [3.1]  Page No [6,7]  Line [2] |
| Explain why ensemble models consistently outperformed other models. | | he superior performance of ensemble models like random forest, gradient boosting, and XGBoost over DL models such as CNN, MobileNet and swin transformer can be attributed to their ability to aggregate multiple weak learners, which enhances generalization and reduces overfitting by enhancing generalization. The models produced almost perfect metrics, having an accuracy of 0.9968,  an F1 score of 0.9981 and an AUC-ROC of 1.00 with stratified 5-fold cross validation. These models are capable of capturing the complex patterns with the use of different decision trees and iterative error correction than simple models such as logistic regression (accuracy of 0.9463) and GNB (accuracy of 0.9150), the CNN model (0.9838 accuracy, 0.9927 F1-score) slightly outperformed the ensembles, while MobileNet (0.9762 accuracy) and swin transformer (0.9421 accuracy) were  computationally intense and overfitting. | Section No [5]  Page no [19,20]  Line No [4] |
| Discuss the implications of the varying training times for practical deployment. | | The varying training times of ML models for lung cancer prediction significantly impact their practical deployment. Models like XGBoost and CNN, with training times of around 15.00 seconds, are more feasible for real time clinical applications due to their efficiency, whereas SVM’s extensive training time (SVM 510.02 seconds) may hinder its use in time-sensitive settings. Balancing high accuracy with shorter training durations is crucial for integrating these models into resource-constrained healthcare environments. | Section No [5] Page No [20]  Line No [6] |
| 2 | better to explain data that they used. And correlation among those parameters. If the target is to improve cancer identification, I think higher accuracy may be achieved by using images. | | Added the Dataset description as a table. | Section No [3.1]  Page no [7]  Line no [1] |
| 3 | Properly cite the dataset. Justify the dataset selection. Include detailed dataset description. | | Added the Dataset description as a table.  The dataset was chosen for its focus on numerical attributes, enabling cost-effective lung cancer prediction without relying on expensive imaging techniques. | Section No [3.1]  Page no [7]  Line no [1]  Section No [3.1]  Page No [6]  Line [2] |
| Enhance literature review | | The related work discusses recent developments in ML and DL for predicting  lung cancer and the application of such approaches using numerical data. It also examines how the hyperparameter optimization strategy and cross-validation  methods are utilized to ensure the models’ robustness and reliability.  ML and DL methods have gained rapid development in lung cancer prediction  over the last decade [3]. SVM alongside random forest, can surpass traditional learning techniques in lung cancer prediction, when processed with advanced sets of features [4]. Ensemble models have generated accurate and stable results compared to traditional models in the literature. Ensemble models such as GNB,  SVM, Logistic Regression, Random Forest, Gradient Boosting and XGBoost were accurate in classifying cancer patients using features such as age, smoking history and symptoms [5, 6]. DL models such as Convolutional Neural Networks (CNNs), MobileNet and Swin Transformer performed well in extracting com-  plex patterns, whereby CNNs only capture hierarchical features, MobileNet is capable of capturing low-resource-efficient tasks, and Swin Transformer models long-range dependencies using an attention mechanism [1]. The transfer learning also enhances the performance, particularly where there is insufficient training data available [7]. Hyperparameter tuning techniques, including Bayesian optimization, have been used to optimize these models by sensitivity adjustment of parameters that improve on model generalization and avoid overfitting (e.g.,learning rates, dropout rates). Besides, cross-validation procedures such as 5-fold, stratified 5-fold and Leave-One-Out (LOO) are needed to gain information on the robustness of a model to ensure reliable and precise clinical decision support systems [8].  Classification models such as Rule-Based, Decision Tree, Naive Bayes and Artificial Neural Network (ANN) are used to detect lung cancer using a massive volume of data [1]. In their study, the model was designed using features such as age, sex, wheezing, shortness of breath, and shoulder, chest, and arm. A comparative analysis revealed that SVM (95.56% accuracy) is accurate at cancer detection, outperforming CNN and K-Nearest Neighbor (KNN) (92.11% and 88.40% accuracy) models, for early lung cancer diagnosis using the UCI dataset, including patients who received a lung cancer diagnosis [5].  Radiomics implies automatic extraction of medical image-based quantitative features, which is widely used for lesion classification applications. Different imaging techniques like CT are being used in lesion investigation, where DL models are employed for automatic extraction of medical image-based features.  A study reviewed the primary methods that classify nodules and predict lung cancer by analyzing CT imaging data [5]. The results revealed that CNNs trained with sufficient data performed best, with an Area Under Curve (AUC) of 0.90; after, it is required to pay careful attention to data limitations present in the validation and training datasets during system performance assessments. The  ensemble model’s prediction capability was compared with ResNet-50, VGG-16, and EfficientNet-B5 DL models with automated feature extraction of histopathological images using a U-Net model [6]. The results revealed that the ensemble model performed best with an accuracy of 0.99, followed by the EfficientNet-B5 with an accuracy of 0.97.  Lung cancer incidence rates of males and females across ten European countries were evaluated using support vector regression (SVR), backpropagation and Long-Short Term Memory networks (LSTM) before lung cancer prediction [7].  Effective assessment metrics, including mean square error (MSE), coefficient of determination (R2) and explained variance (EV) scores, were used for the results evaluation, where SVR recorded the best performance and LSTM recorded the lowest performance.  The prospect of incorporating various features and refined hyperparameters to achieve diagnostic precision in non-small cell lung cancer (NSCLC) precisely and small cell lung cancer (SCLC) is studied in literature [8]. Hybrid feature extraction of grey-level co-occurrence matrix (GLCM), Haralick and autoencoder features, and optimized machine learning models were used to develop accurate  lung cancer detection models. The study results showed that SVM, radial basis functions (RBF) and SVM gaussian models with hybrid features and SVM polynomial with single Haralick features improved the accuracy of the models.  Boosting models like XGBoost and LightGBM are viable predictive models exhibiting superior performance when compared with AdaBoost, Logistic Regression and SVM [9]. The analysis revealed that XGBoost consistently outperformed the other models in terms of accuracy, sensitivity, specificity and F1 score, achieving 97.50%, 96.80%, 98%, and 97.50%. LightGBM also showed strong results, remaining as a potential alternative.  However, there are still limitations and challenges in the previous research to predict lung cancer. Homogeneous or small datasets are hardly representative of the diverse populations of patients that make the models less generalizable [1]. The excessive focus on accuracy as an essential indicator might ignore other vital clinical indicators such as sensitivity and specificity in the predisposition to unreliable forecasts. Transformers exhibit computational complexity, requiring large resources, an aspect that increases their limitation to low-resource environments [7, 10]. Features such as poor cross-validation strategies and inability to interpret the model weaken the trust placed in a given model by physicians, particularly due to the risk of overfitting [9]. Real-world applications are fur-  ther complicated by poor integration into clinical workflows and adjustable data  preprocessing, including processing of missing values or normalization [11]. | Section No [2]  Page No [2-4]  Line [1-8]. |
| Describe in detail how cross-validation was used in both hyperparameter tuning and final model evaluation | | Both hyperparameter optimization and evaluation of the final model were done using stratified 5-Fold cross-validation with k=5. The dataset was split into stratified into five folds (YES/NO Lung cancer) while mitigating the imbalance issue. Each fold was used as a validation set once, and the remaining four as training to tune the hyperparameters using Bayesian optimization to maximize the accuracy. Stratified 5-fold cross validation with optimized hyperparameters was used to evaluate the final models. | Section No [3.5]  Page No [12]  Line No [2] |
| Justify the choice of k to 5 | | The choice of k=5 in stratified 5-Fold cross-validation is due to its compromise between reliability and efficiency. It allows dividing the data into 80 training and 20 validation per fold, which is enough to train on, and it gives stable performance estimates with low variance compared to k=3 and Holdout [6]. k=5 is computationally efficient in comparison to k=10 and LOOCV [7]. In the case of models such as SVM (510.02s training time) and appropriate to the size of the dataset, as demonstrated by XGBoost rapid 15.41s training time and stable accuracy of 0.9968. Stratification preserves the proportion of classes, which is best in terms of reliability with imbalanced data thus, it is the best option in this study | Section No [5]  Page [19],  Line No [3] |