Feature Engineering with Large Language Models Improves Solitary Pulmonary Nodule Malignancy Classification with Machine Learning

Ioannis D. Apostolopoulos
University of Thessaly
Department of Energy Systems
Larissa, Greece
ece7216@upnet.gr

Nikolaos D. Papathanasiou
University Hospital of Patras
Department of Nuclear Medicine
Patras, Greece
nikopapath@upatras.gr

Dimitrios Apostolopoulos
University Hospital of Patras
Department of Nuclear Medicine
Patras, Greece
dimap@med.upatras.gr

Nikolaos Papandrianos
University of Thessaly
Department of Energy Systems
Larissa, Greece

Elpiniki I. Papageorgiou
University of Thessaly
Department of Energy Systems
Larissa, Greece

2025 16th International Conference on Information, Intelligence, Systems & Applications (IISA) | 979-8-3315-5636-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/IISA66859.2025.11311233

***Abstract*— Accurate classification of solitary pulmonary nodules (SPNs) as benign or malignant can be critical for lung cancer diagnosis and timely treatment. Traditional machine learning approaches rely on hand-crafted features for the classification task. This study explores the use of large language models (LLMs) for automated feature engineering to enhance SPN malignancy classification using a Random Forest classifier. A baseline dataset containing five standard radiological features was used to train a Random Forest model. Multiple LLMs, including GPT-4.0, Gemini, and others, were prompted to propose up to five new, clinically plausible features derived from or related to the original features. The suggested extra features were incorporated into new feature sets and evaluated using accuracy, sensitivity, and specificity metrics. All LLM-enhanced feature sets improved the classifier (88.52% accuracy), with the best results achieved using features proposed by GPT-4.0, reaching 94.64% accuracy, 96.16% sensitivity, and 93.54% specificity. Recurrent high-impact features included the SUVmax-to-Diameter Ratio, Margin Irregularity Index, and Nodule Growth Rate. LLMs show significant promise for automated feature engineering in clinical machine learning. Their ability to generate medically interpretable and performance-enhancing features can accelerate model development and improve diagnostic accuracy in lung cancer screening.**

Keywords—Large Language Models; Feature Engineering; Machine Learning; Solitary Pulmonary Nodules

I. INTRODUCTION

Lung cancer is a major global health challenge, representing a significant cause of morbidity and mortality. [1], [2]. The early detection and characterization of solitary pulmonary nodules (SPNs), which are small, rounded opacities typically found incidentally during chest imaging [3] is a significant challenge. SPNs are radiographically defined as isolated lesions smaller than 3 cm in diameter that are surrounded by lung parenchyma and not associated with atelectasis or lymphadenopathy [4]. Timely and accurate discrimination between benign and malignant SPNs is essential for early intervention.

Traditional approaches for SPN classification rely on a combination of imaging features and clinical judgment. In recent years, machine learning (ML) has emerged as a promising tool to enhance diagnostic accuracy by modeling complex, nonlinear interactions between radiological features [5], [6], [7], [8], [9]. Among ML methods, tree-based models like Random Forests (RF) and Gradient Boosted Trees (GBT) have demonstrated effectiveness due to their robustness and ability to handle high-dimensional, non-parametric data. These models are trained on a set of tabular features such as nodule diameter, location, margin type, and SUVmax (maximum standardized uptake value on PET-CT), which have shown relevance in malignancy risk estimation [10], [11].

However, the performance of ML models often hinges on the quality of the input features. Feature engineering, which transforms and creates new input variables, might significantly influence model effectiveness. Conventional feature engineering typically requires substantial domain expertise and manual effort. Recent advancements in artificial intelligence (AI), specifically large language models (LLMs), offer a novel approach to automated feature engineering [12].

This paper investigates the potential of LLMs such as GPT-4 and Gemini in automating feature engineering for SPN classification. Recent studies demonstrated that similar models can offer major contribution to feature engineering tasks [13], [14].

II. MATERIALS AND METHODS

A. Dataset

Between 2018 and 2022, over 800 PET/CT scans showing confirmed lung nodules were thoroughly reviewed at the Laboratory of Nuclear Medicine in the University Hospital of Patras, Greece, to identify eligible study participants. Scans without SPNs or showing nodules smaller than 0.6 cm or larger than 3 cm were excluded. 456 scans were suitable for analysis. The clinical team extracted detailed SPN characteristics from these, including size, type, edge definition, and SUVmax uptake levels.

SPN malignancy was determined using three approaches: (a) biopsy findings, (b) follow-up, or (c) expert judgment based on imaging and clinical indicators such as patient age, smoking history (including secondhand exposure), occupational risks, and past lung conditions.

The average participant age was 66, with a gender distribution of 69% male and 31% female. Of the nodules analyzed, 51% were classified as malignant and 49% as benign. Strict protocols were followed to ensure ethical data handling. PET, CT, and clinical data were collected under rigorous standards, with all identifying DICOM information removed immediately. Patient records were anonymized using coded identifiers. These practices adhered to the Declaration of Helsinki, prioritizing patient privacy and research integrity.

B. Classification model

A Random Forest (RF) classifier was employed as the predictive model for SPN malignancy classification. Random Forests are ensemble learning models that build multiple decision trees and aggregate their outputs to improve generalization and reduce overfitting [15].They are particularly effective for handling structured, tabular data and allow for evaluating feature importance.

The model was trained and validated using the base dataset consisting of five features (Diameter, SUVmax, Nodule Type, Nodule Location, and Nodule Margins) and extended versions including LLM-suggested features. Model performance was evaluated using three standard metrics: accuracy (ACC), sensitivity (SEN), and specificity (SPE). We used a random train-test split (70-30) to compute the performance metrics.

C. Language Models

Large Language Models (LLMs) are promising tools for reasoning, content generation, and problem-solving across many domains, including biomedical informatics and clinical decision support [16], [17]. Trained on trillions of tokens, these models use transformer-based architectures that enable contextual understanding and natural language reasoning.

Recent studies have demonstrated the capability of LLMs to perform complex tasks such as medical summarization, question answering, and even radiological interpretation [16], [18], [19]One emerging application is the use of LLMs in data preprocessing and feature engineering, where models are prompted to derive new variables or transformations based on structured or unstructured input.

In this study, each LLM received a structured prompt (Table I), positioning it as an expert in machine learning, feature engineering, and medical data analysis. LLMs were allowed to suggest plausible transformations or conceptual features, with the constraint that they be medically interpretable and potentially measurable in practice.

The LLMs used in this study include:

- **GPT-3.5 / GPT-4.0 / GPT-4 Turbo**: Developed by OpenAI, these models are competent at multitask reasoning and contextual feature inference.
- **o1-mini / Copilot**: Smaller-scale yet optimized LLMs built for task-specific completions.
- **Llama 3.1 Sonar, DeepSeek V3, Liquid LFM 40B, Gemini**: High-capacity open-source or proprietary models designed for extensive prompt handling and multi-turn reasoning.

Each model was evaluated based on its suggested features' originality, plausibility, and impact.

TABLE I. PROMPT TO THE LANGUAGE MODELS

You are an expert in machine learning, feature engineering, and medical data analysis. Your goal is to enhance a classification model that predicts whether solitary pulmonary nodules (SPNs) are Benign or Malignant based on input features. The current features used in the model and their potential values are:

1 Input Features:
1. Diameter: 0-3 cm
2. SUVmax: Continuous variable
3. Nodule Type: {Solid, Semi-solid, Ground-glass}
4. Nodule Location: {Left Lower Lobe (LLL), Lingula, Middle, Right Upper Lobe (RUL), Right Lower Lobe (RLL)}
5. Nodule Margins: {Well-defined, Lobulated, Spiculated, Ill-defined}

2 Constraints:
- Other available features such as Age, Gender, and GLU are excluded from the analysis due to expert opinion deeming them insignificant, but you may use them if you believe they are useful.

3 Task:

Propose up to 5 new features that could improve the predictive performance of the model. These features can be derived from the existing features through combinations or transformations, or they can be entirely new features that could be hypothetically collected. Your suggestions should be scientifically plausible and relevant to SPN malignancy prediction.

4 Guidelines:
1. Suggest combinations of existing features that may reveal interactions or higher-order relationships relevant to malignancy.
2. Propose new, plausible features if they would likely be informative and could be measured in practice.
3. Ensure the suggested features are medically interpretable and feasible for use in a clinical setting.

5 Model Context:
- The ML algorithm used for this task is a tree-based classifier (e.g., Random Forest or Gradient Boosted Trees).
- The two classes to predict are Benign SPN and Malignant SPN.

6 Output Format:

For each suggested feature:
1. Name of the feature
2. Description of how it is derived (if applicable)

Explanation of why it might be predictive of malignancy

III. RESULTS

A. Suggested features

The LLMs generated diverse novel features derived from transformations, interactions, or higher-order relationships among the base variables. Table II presents a comprehensive list of features proposed by each language model.

Key recurring features across multiple models include:

- **Margin Irregularity Index**: A score derived from categorizing and quantifying the spiculation or irregularity in the nodule margin, reflecting malignancy potential.
- **SUVmax-to-Diameter Ratio**: A normalized intensity feature suggesting that metabolic activity

per unit size may be a more decisive indicator of malignancy.

- **Nodule Growth Rate**: Although not directly included in the static dataset, several models proposed temporal derivatives assuming prior imaging, highlighting its clinical relevance.
- **Location-Based Risk Score**: Based on literature trends, a risk index mapping anatomical SPN location to empirical malignancy likelihood.
- **Nodule Density Score / Texture**: Proposed as a quantitative metric derived from CT scan heterogeneity, which can indicate internal structural complexity.

Some models extended feature engineering into more speculative territory:

- **Nodule Morphology Score** (GPT-4): A composite score integrating margin, type, and shape characteristics.
- **Nodule Proximity to Hilum** (Gemini): A spatial feature hypothesized to relate to lymphatic involvement likelihood.
- **SUVmax Variability Over Time** (Liquid LFM 40B): A dynamic feature requiring longitudinal imaging, reflecting progression or regression.

Independent LLMs' repeated suggestions of certain features, especially the Margin Irregularity Index and SUVmax-to-Diameter Ratio, demonstrate both plausibility and clinical resonance. Furthermore, these engineered features produced measurable performance improvements in the subsequent classification model (as detailed in Section III.B) affirms their usefulness.

TABLE II. LLM-SUGGESTED FEATURES

Model	Suggested Features
GPT 3.5	Nodule Shape; Age-Adjusted SUVmax; Nodule Position Index; Nodule Growth Rate; Nodule Texture;
GPT 4.0	Volume-to-SUVmax Ratio; Nodule Morphology Score; Location-Based Risk Score; Diameter-SUVmax Interaction Term; Margin Irregularity Index;
GPT 4.0 Turbo	Nodule Growth Rate; Nodule Density Variation; Margin Irregularity Index; Nodule Volume; SUVmax-to-Diameter Ratio;
o1-mini	Margin Irregularity Index; SUVmax-to-Diameter Ratio; Lobe Malignancy Risk Index; Nodule Growth Rate; Nodule Density Score;
ChatGPT 4	Margin Irregularity Index; SUVmax-to-Diameter Ratio; Lobe Malignancy Risk Index; Nodule Density Transformation; Volume-to-SUVmax Ratio;
Copilot	Nodule Volume; SUVmax-to-Diameter Ratio; Nodule Density; Location-Based Risk Score; Margin Irregularity Index;
Llama 3.1 Sonar Large 128K	Nodule Growth Rate; Margin Irregularity Index; Location-Based Risk Score; Nodule Density Heterogeneity; Nodule-Adjacent Pleural Involvement;
Liquid LFM 40B	Nodule-to-Lobe Volume Ratio; SUVmax-to-Diameter Ratio; Margin Irregularity Index; Nodule Location in Relation to Airways; SUVmax Variability Over Time;
DeepSeek V3	Nodule Density Score; Margin Irregularity Index; Location-Based Risk Score; SUVmax-to-Diameter Ratio; Nodule Growth Rate;
Gemini	SUVmax-to-Diameter Ratio; Nodule Location Proximity to Hilum; Nodule Margin Sharpness; Nodule Type and Location Interation; Nodule Growth Rate;

B. Impact on the classification performance

Table III compares the classification performance of the base input feature set and the augmented feature sets, as suggested by the LLMs. Working with the original feature set yielded an accuracy of 88.52%, with a sensitivity of 91.04% and specificity of 86.68%. Adding new features, regardless of their combinations, increased the performance across most cases. In the best-performing case, extra features by GPT 4.0 were used, achieving the highest accuracy at 94.64%, the highest sensitivity at 96.16%, and the highest specificity at 93.54%. Gemini was another strong performer in feature suggestion, improving the results to an accuracy of 94.33%, sensitivity of 95.94%, and specificity of 93.15%, slightly lower than GPT 4.0. Other models, such as GPT 4.0 Turbo and Copilot, improved, with accuracy of 93.79% and 93.7%, respectively. These models improved sensitivity and specificity, but did not surpass GPT 4.0 in overall performance. Some models, such as Liquid LFM 40B and DeepSeek V3, achieved high sensitivity, but their specificity is lower than others. Similarly, Llama 3.1 exhibited accuracy of 91.36%, sensitivity of 94.02%, and specificity of 89.41%.

TABLE III. CLASSIFICATION METRICS FOR MULTIPLE INPUT FEATURE COMBINATIONS

Input Features	ACC (%)	SEN (%)	SPE (%)
Base	88.52	91.04	86.68
Base + GPT 3.5	90.95	94.34	88.47
Base + GPT 4.0	**94.64**	**96.16**	**93.54**
Base + GPT 4.0 Turbo	93.79	95.62	92.45
Base + o1-mini	93.43	95.62	91.82

Base + ChatGPT 4	93.02	94.98	91.59
Base + Copilot	93.7	95.62	92.29
Base + Llama 3.1 Sonar Large 128K	91.36	94.02	89.41
Base + Liquid LFM 40B	90	96.05	85.59
Base + DeepSeek V3	90.32	94.13	87.54
Base + Gemini	94.33	95.94	93.15

IV. Discussion and Conclusions

This study explored using LLMs for automated feature engineering in classifying SPNs using machine learning. We demonstrated that model performance can be significantly enhanced beyond achievable using standard radiological features alone.

The best-performing feature augmentation came from GPT-4.0, whose suggestions led to a classification accuracy of 94.64%, with sensitivity and specificity exceeding 93%, a notable improvement over the baseline accuracy of 88.52%. Other LLMs, including Gemini and Copilot, also showed strong performance, indicating that this approach is not model-specific but generalizable across various LLM architectures.

LLMs repeatedly proposed several features, such as SUVmax-to-Diameter Ratio, Margin Irregularity Index, and Nodule Growth Rate, represent already established clinical heuristics. This highlights the ability of LLMs to mimic human expert reasoning and formalize it into features. Moreover, the emergence of composite indices (e.g., Nodule Morphology Score, Location-Based Risk Score) suggests that LLMs can generalize from prior training to generate medically plausible constructs, even when not explicitly observed during pretraining.

These findings underline the promise of integrating LLMs into the machine learning pipeline, particularly in the feature engineering phase, which remains one of the most labor-intensive and expertise-driven components of model development. Despite encouraging results, this study has limitations. First, we did not evaluate multicollinearity or auto-correlation among the engineered features, which could lead to overfitting or unstable model behavior in real-world deployment. Second, the data was limited to a single internal dataset without external validation. Therefore, the generalizability of the LLM-generated features across institutions or imaging protocols remains to be confirmed.

References

[1] L. A. Torre, R. L. Siegel, and A. Jemal, "Lung Cancer Statistics," in *Lung Cancer and Personalized Medicine*, vol. 893, A. Ahmad and S. Gadgeel, Eds., Cham: Springer International Publishing, 2016, pp. 1–19. doi: 10.1007/978-3-319-24223-1_1.

[2] B. C. Bade and C. S. D. Cruz, "Lung cancer 2020: epidemiology, etiology, and prevention," *Clinics in Chest Medicine*, vol. 41, no. 1, pp. 1–24, 2020.

[3] J. P. Ko, B. Bagga, E. Gozansky, and W. H. Moore, "Solitary Pulmonary Nodule Evaluation: Pearls and Pitfalls," *Seminars in Ultrasound, CT and MRI*, vol. 43, no. 3, pp. 230–245, Jun. 2022, doi: 10.1053/j.sult.2022.01.006.

[4] A. Cruickshank, G. Stieler, and F. Ameer, "Evaluation of the solitary pulmonary nodule," *Internal Medicine Journal*, vol. 49, no. 3, pp. 306–315, Mar. 2019, doi: 10.1111/imj.14219.

[5] I. D. Apostolopoulos *et al.*, "Automatic classification of solitary pulmonary nodules in PET/CT imaging employing transfer learning techniques," *Med Biol Eng Comput*, vol. 59, no. 6, pp. 1299–1310, Jun. 2021, doi: 10.1007/s11517-021-02378-y.

[6] Y.-J. Yu-Jen Chen, K.-L. Hua, C.-H. Hsu, W.-H. Cheng, and S. C. Hidayati, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," *OTT*, p. 2015, Aug. 2015, doi: 10.2147/OTT.S80733.

[7] Y. S. Salihoğlu *et al.*, "Diagnostic Performance of Machine Learning Models Based on 18F-FDG PET/CT Radiomic Features in the Classification of Solitary Pulmonary Nodules," *Mirt*, vol. 31, no. 2, pp. 82–88, Jun. 2022, doi: 10.4274/mirt.galenos.2021.43760.

[8] H. Wang *et al.*, "A machine learning-based PET/CT model for automatic diagnosis of early-stage lung cancer," *Front. Oncol.*, vol. 13, p. 1192908, Sep. 2023, doi: 10.3389/fonc.2023.1192908.

[9] R. P. Shah *et al.*, "Machine Learning Radiomics Model for Early Identification of Small-Cell Lung Cancer on Computed Tomography Scans," *JCO Clinical Cancer Informatics*, no. 5, pp. 746–757, Dec. 2021, doi: 10.1200/CCI.21.00021.

[10] I. D. Apostolopoulos, N. D. Papathanasiou, D. J. Apostolopoulos, N. D. Papandrianos, and E. I. Papageorgiou, "Investigating the Agreement with Human Readers and Generalisation Capabilities of a Transfer Learning Approach for Predicting the Malignancy of Solitary Pulmonary Nodules in CT Screening," in *2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Chania Crete, Greece: IEEE, Jul. 2024, pp. 1–6. doi: 10.1109/IISA62523.2024.10786628.

[11] I. D. Apostolopoulos, N. D. Papathanasiou, D. J. Apostolopoulos, E. I. Papageorgiou, and N. Papandrianos, "A machine learning approach for determining solitary pulmonary nodule malignancy in patients undergoing PET/CT examination," *Multimed Tools Appl*, Mar. 2025, doi: 10.1007/s11042-025-20737-x.

[12] S. Malberg, E. Mosca, and G. Groh, "FELIX: Automatic and Interpretable Feature Engineering Using LLMs," in *Machine Learning and Knowledge Discovery in Databases. Research Track*, vol. 14944, A. Bifet, J. Davis, T. Krilavičius, M. Kull, E. Ntoutsi, and I. Žliobaitė, Eds., in Lecture Notes in Computer Science, vol. 14944. , Cham: Springer Nature Switzerland, 2024, pp. 230–246. doi: 10.1007/978-3-031-70359-1_14.

[13] T. Isomura, R. Shimizu, and M. Goto, "LLMOverTab: Tabular data augmentation with language model-driven oversampling," *Expert Systems with Applications*, vol. 264, p. 125852, Mar. 2025, doi: 10.1016/j.eswa.2024.125852.

[14] A. M. Kashyap, D. Rao, M. R. Boland, L. Shen, and C. Callison-Burch, "Predicting explainable dementia types with LLM-aided feature engineering," *Bioinformatics*, vol. 41, no. 4, p. btaf156, Mar. 2025, doi: 10.1093/bioinformatics/btaf156.

[15] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, Que., Canada: IEEE Comput. Soc. Press, 1995, pp. 278–282. doi: 10.1109/ICDAR.1995.598994.

[16] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nat Med*, vol. 29, no. 8, pp. 1930–1940, Aug. 2023, doi: 10.1038/s41591-023-02448-8.

[17] Y. Chang *et al.*, "A Survey on Evaluation of Large Language Models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, Jun. 2024, doi: 10.1145/3641289.

[18] N. H. Shah, D. Entwistle, and M. A. Pfeffer, "Creation and Adoption of Large Language Models in Medicine," *JAMA*, vol. 330, no. 9, p. 866, Sep. 2023, doi: 10.1001/jama.2023.14217.

[19] J. Clusmann *et al.*, "The future landscape of large language models in medicine," *Commun Med*, vol. 3, no. 1, p. 141, Oct. 2023, doi: 10.1038/s43856-023-00370-1.