Lung Cancer Canada
Give A Breath Research Award Committee

Subject: Letter of Intent for the Give A Breath Research Award

Dear Members of the Give A Breath Research Award Committee,

I am writing to express my intent to apply for the Give A Breath Research Award in support of my research project, *"AI-Enhanced Liquid Biopsy for Monitoring Treatment Resistance in Advanced Lung Cancer."* This research aligns with the award's mission to address urgent needs in advanced lung cancer, particularly in patients who progress beyond first-line therapy. By advancing non-invasive monitoring strategies, our work seeks to improve diagnostic precision, predict treatment resistance, and optimize patient outcomes for Canadians living with stage III or IV lung cancer.

Our project proposes the development and validation of an artificial intelligence (AI)-powered liquid biopsy platform that integrates circulating tumor DNA (ctDNA) and proteomic biomarkers. By focusing on clinically relevant mutations (EGFR, KRAS, ALK, TP53) alongside proteomic signatures, we will generate a compact biomarker panel that can accurately identify resistance and disease progression. Building on our recent publication in *iScience* (2023), which demonstrated the power of deep learning to improve biomarker discovery, this proof-of-concept project will leverage public datasets and collaborative expertise to deliver actionable insights that can be readily applied in clinical oncology practice.

This research is a collaborative effort with Dr. Rosalyn Juergens, thoracic oncologist and expert in liquid biopsy applications in lung cancer, and Dr. Zhiyong Zhang of Stanford University for AI in Healthcare, a leader in scalable deep learning and biomarker integration. Their combined expertise in clinical oncology and AI-driven healthcare innovation will be essential to ensuring both feasibility and clinical relevance.

By targeting the critical gap in post-first-line therapy management, this project will drive a transformative shift in how we monitor advanced lung cancer, ultimately reducing unnecessary biopsies, guiding treatment selection, and improving survival and quality of life for patients.

I sincerely appreciate your consideration of this application and look forward to contributing to the advancement of lung cancer research and patient care. Please do not hesitate to contact me if you require further information.

Sincerely,

Fei Geng

Background and Rationale

Advanced lung cancer (stage III/IV) patients who progress after first-line therapy face limited treatment options and poor outcomes[1]. Monitoring depends largely on invasive tissue biopsies, which are impractical to repeat. Liquid biopsy offers a non-invasive alternative, but its reliability is limited by low ctDNA abundance and biological variability[2]. Artificial intelligence (AI) provides an opportunity to overcome these barriers by detecting complex biomarker patterns that predict resistance and progression[3–5]. Our team has pioneered this approach. In 2023, the Geng Research Group published in *iScience* a deep learning feature-extraction framework that improved biomarker discovery and classification accuracy compared with conventional methods[3]. This work demonstrated robust class separation and interpretable biomarker identification. Building on this foundation, we propose to integrate ctDNA mutation profiles, protein biomarkers, and clinical metadata to develop an AI-powered tool for detecting resistance and progression in advanced lung cancer.

Hypothesis and Objectives

We hypothesize that an AI framework (Fig. 1) integrating liquid biopsy multi-omics will identify predictive signatures of therapy resistance and disease progression in advanced lung cancer patients beyond first-line therapy. Our objectives are: (1) to develop and validate a multi-input deep learning model incorporating ctDNA, protein biomarkers, and clinical metadata; (2) to demonstrate that the model improves classification of progression versus response compared with current monitoring approaches; (3) to evaluate predictive biomarkers for correlation with clinical outcomes such as progression-free survival.

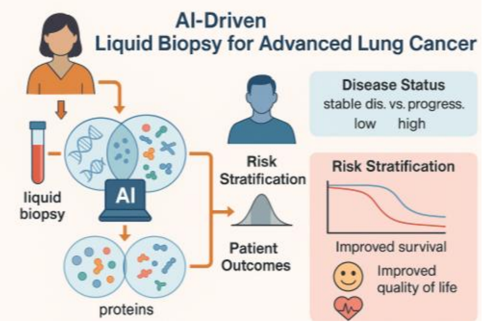

Figure 1. AI-enhanced liquid biopsy for advanced lung cancer. Liquid biopsy samples of ctDNA and proteins are analyzed by an AI model to stratify risk and distinguish stable disease from progression, enabling oncologists to optimize treatment decisions. This approach supports earlier intervention, improved survival, and better quality of life for patients.

Feasibility Statement

This project is designed to be feasible within the funding limit by leveraging publicly accessible, well-curated datasets rather than initiating a costly prospective collection. By narrowing ctDNA analysis to the most clinically relevant mutations (EGFR, KRAS, ALK, TP53) and integrating these with available proteomic and clinical data, we ensure that the scope is realistic and achievable. Existing institutional resources and expertise in AI modeling and biomarker analysis further support feasibility. The proposed work is therefore appropriately scaled for a one-year, proof-of-concept study and will provide critical preliminary data to enable larger, externally funded trials.

Team and Roles

This project will be conducted by a multidisciplinary team with expertise spanning oncology, artificial intelligence, and biomarker research.

Dr. Fei Geng, PhD (Principal Investigator): Associate Professor in Faculty of Engineering at McMaster University. Dr. Geng has extensive experience in cancer diagnostics[6], biomarker discovery[7], and AI applications[3,8] for lung cancer. He will oversee project design, AI framework development, and integration of ctDNA/proteomic data.

Dr. Rosalyn Juergens, MD, PhD (Clinical Oncology Lead): Thoracic oncologist at Hamilton Health Sciences and Professor of Oncology at McMaster University. Dr. Juergens will guide clinical relevance[9,10], ensuring that the biomarker and AI findings are aligned with patient needs and treatment decision-making. She will provide access to clinical expertise and validation opportunities in lung cancer management.

Dr. Zhiyong Zhang, PhD (AI and Computing Lead): Senior Scientist in AI based Healthcare at Stanford University. Dr. Zhang brings expertise in scalable deep learning architectures, GPU-accelerated computing, and integration of large-scale biomarker datasets[11,12]. He will advise on computational efficiency and optimization of the multi-modal AI framework, ensuring that the platform can be scaled to larger cohorts in future studies.

Methodology

1. Study Design: This is a translational research project combining AI model development with publicly accessible, well-curated clinical and biomarker datasets. Instead of initiating a costly prospective collection, we will use recommended publicly available datasets that include ctDNA mutation data, proteomic panels, and associated clinical outcomes in advanced NSCLC patients.

2. Focused Proof-of-Concept Scope: To ensure feasibility, ctDNA profiling will be restricted to the most clinically relevant mutations (e.g. EGFR, KRAS, ALK, and TP53). These are frequently altered in NSCLC and directly inform clinical decision-making. This narrower scope is cost-effective while remaining sufficient to demonstrate proof-of-principle. Positive results from this targeted analysis will justify expansion to broader panels and prospective cohorts in future, larger-scale studies.

3. Biomarker Analysis: Public datasets will provide ctDNA profiles for key driver mutations and proteomic measurements for circulating biomarkers. Where available, additional clinical metadata (e.g. demographics, prior treatments, outcomes) will be included. This will ensure robust integration and correlation with disease progression.

4. AI Integration: Building upon our 2023 *iScience* methodology, we will adapt the feature-extraction framework to multi-modal inputs, training the AI to combine ctDNA and proteomic patterns to predict progression. Validation will use cross-validation on public datasets, with interpretability achieved to identify the most predictive biomarkers. This compact biomarker set will form the foundation of a clinically implementable assay.

Timeline and Stages

The proposed research will be completed within the one-year duration of the award, divided into three stages: **Stage 1 (Months 1–4):** Dataset acquisition, ethics confirmation for secondary use, and preprocessing pipeline adaptation. **Stage 2 (Months 5–8):** Initial AI model training using ctDNA and proteomic data focused on EGFR, KRAS, ALK, and TP53. **Stage 3 (Months 9–12):** Algorithm refinement with incorporation of clinical metadata and feature attribution analysis.

Expected Outcomes

This project will produce proof-of-concept validation of an AI-enhanced liquid biopsy platform for advanced lung cancer. Expected outcomes include a compact biomarker panel of ctDNA and proteins with predictive utility, reduced reliance on repeat biopsies, earlier detection of progression, and improved clinical decision-making beyond first-line therapy. Ultimately, this tool could enable oncologists to intervene earlier, select more effective second-line treatments, and improve survival and quality of life for Canadian patients.

1. Langer, C. J., Mok, T. & Postmus, P. E. Targeted agents in the third-/fourth-line treatment of patients with advanced (stage III/IV) non-small cell lung cancer (NSCLC). *Cancer Treat Rev* 39, 252–260 (2013).

2. Ma, L. *et al.* Liquid biopsy in cancer current: status, challenges and future prospects. *Signal Transduct Target Ther* 9, 1–36 (2024).

3. Tsai, Y. *et al.* Targeted deep learning classification and feature extraction for clinical diagnosis. *iScience* 26, (2023).

4. Alum, E. U. AI-driven biomarker discovery: enhancing precision in cancer diagnosis and prognosis. *Discover Oncology* 16, 313 (2025).

5. Casagrande, G. M. S., Silva, M. de O., Reis, R. M. & Leal, L. F. Liquid Biopsy for Lung Cancer: Up-to-Date and Perspectives for Screening Programs. *Int J Mol Sci* 24, 2505 (2023).

6. Kong, Q. *et al.* Analysis of the susceptibility of lung cancer patients to SARS-CoV-2 infection. *Mol Cancer* 19, 80 (2020).

7. GENG, F., SHI, B. Z., YUAN, Y. F. & WU, X. Z. The expression of core fucosylated E-cadherin in cancer cells and lung cancer patients: prognostic implications. *Cell Res* 14, 423–433 (2004).

8. Yu, W., Bai, Y., Raha, A., Su, Z. & Geng, F. Integrative In Silico Investigation Reveals the Host-Virus Interactions in Repurposed Drugs Against SARS-CoV-2. *Frontiers in Bioinformatics* 0, 81 (2022).

9. Melosky, B. *et al.* Modern era systemic therapies: Expanding concepts of cure in early and locally advanced non-small cell lung cancer. *Int J Cancer* 155, 963–978 (2024).

10. Breadner, D. *et al.* Implementation of Liquid Biopsy in Non-Small-Cell Lung Cancer: An Ontario Perspective. *Current Oncology 2024, Vol. 31, Pages 6017-6031* 31, 6017–6031 (2024).

11. Plasser, F. *et al.* COLUMBUS—An Efficient and General Program Package for Ground and Excited State Computations Including Spin–Orbit Couplings and Dynamics. *J Phys Chem A* (2025) doi:10.1021/ACS.JPCA.5C02047.

12. Lischka, H. *et al.* The generality of the GUGA MRCI approach in COLUMBUS for treating complex quantum chemistry. *J Chem Phys* 152, (2020).

This project is positioned to exert a sustained and powerful influence on both lung cancer research and clinical practice. Advanced and metastatic lung cancer remains the leading cause of cancer-related mortality in Canada, with more than 28,000 new cases annually and five-year survival rates below 20%. The majority of patients with non–small cell lung cancer (NSCLC) ultimately progress after first-line therapy—whether chemotherapy, immunotherapy, or targeted therapy—and face poor outcomes. In routine practice, disease monitoring is heavily reliant on serial imaging and, when feasible, tissue biopsies. These approaches are invasive, expensive, and often inadequate for capturing dynamic tumor evolution such as emerging resistance mutations or molecular heterogeneity. Our project directly addresses these gaps by **developing an AI-powered liquid biopsy platform** capable of integrating ctDNA and proteomic biomarkers to provide a non-invasive, reliable method for monitoring disease progression and treatment resistance.

The immediate benefit of this research will be to **improve patient quality of life and clinical decision-making**. Patients will be spared from repeated tissue biopsies, reducing procedural pain, pneumothorax risk, and delays in care. By **identifying resistance earlier**, such as detection of EGFR T790M, KRAS G12C, or ALK fusion alterations, **oncologists will be able to act sooner**, discontinuing ineffective therapies, minimizing unnecessary toxicity, and transitioning patients to approved second-line treatments (e.g., osimertinib for EGFR T790M, sotorasib for KRAS G12C, lorlatinib for ALK). This proactive approach will allow earlier therapeutic intervention, directly improving treatment efficacy and extending survival. Families and patients will also benefit from clearer guidance, fewer care disruptions, and greater confidence in treatment planning.

In the short- to medium-term, the project will promote a major advancement in lung cancer research by accelerating the translation of biomarker findings into actionable clinical outcomes. By focusing on a compact and clinically relevant panel of ctDNA mutations (EGFR, KRAS, ALK, TP53) alongside proteomic markers of disease burden and systemic inflammation, the study will **generate data that can be directly applied in Canadian oncology clinics**. This approach supports practical adoption by **aligning with existing treatment algorithms, response assessments, and the need for rapid turnaround in clinical decision-making**. These advances will optimize patient care pathways, improve treatment stratification, reduce unnecessary toxicity, and support the personalization of therapy for patients with advanced lung cancer.

The long-term implications are equally significant. Demonstrating feasibility in this pilot phase will catalyze larger multi-centre trials, positioning Canada at the forefront of precision oncology. Ultimately, **national implementation** of **AI-enhanced liquid biopsy platforms** could reduce mortality by enabling earlier interventions and lowering the incidence of advanced presentations through improved monitoring and proactive treatment adjustment. This initiative will set new standards for patient-centered care, ensuring that knowledge gained from cutting-edge science is translated rapidly into outcomes that matter most: optimized patient care, improved survival, enhanced quality of life, and reduced burden of lung cancer in Canada.

Lung cancer is the leading cause of cancer-related death in Canada, with more people dying from lung cancer each year than breast, prostate, and colorectal cancers combined. Unfortunately, most patients with advanced lung cancer eventually see their disease return or worsen after their first round of treatment. When this happens, doctors face difficult decisions about when to change therapies and which treatments are likely to work best. Today, these decisions often rely on repeat tissue biopsies and frequent scans. Tissue biopsies can be painful, carry medical risks, and are not always possible. Scans, while helpful, do not always show the full picture of how a cancer is changing at the molecular level. This leaves patients and families living with uncertainty and sometimes results in delays in receiving the right treatment.

This project aims to change that by using a simple blood test—often called "liquid biopsy"—combined with artificial intelligence (AI) to monitor lung cancer more effectively. A liquid biopsy looks for tiny pieces of DNA and proteins from the tumor that are circulating in a patient's blood. By analyzing these markers, doctors will learn important information about whether cancer is responding to treatment or developing resistance. However, because these signals are often very small and complex, they can be hard to interpret using traditional methods.

Our research group has developed advanced AI technology that can find patterns in this type of data that would otherwise go unnoticed. By training the AI system to recognize key genetic changes (such as those in EGFR, KRAS, ALK, and TP53), along with protein markers in the blood, we can create a more accurate and reliable test. This test will give oncologists timely information about whether a patient's treatment is still working or needs to be changed.

The potential benefits for patients are significant. With this approach, patients may avoid repeated invasive biopsies, reducing pain, anxiety, and medical risk. They may also spend less time waiting for answers, since a blood test can be performed more quickly than a biopsy. Most importantly, earlier detection of treatment resistance could allow doctors to switch therapies sooner, improving survival and maintaining quality of life. Families will benefit from clearer guidance and fewer disruptions in care.

Although this is a pilot project, it has the potential to make a lasting impact. By showing that AI can improve how liquid biopsy tests are interpreted, this study will lay the foundation for larger trials in Canada and eventually new standards of care. If successful, this work will help ensure that patients with advanced lung cancer receive the right treatment at the right time, with less burden and greater peace of mind.

This project is supported by a team of experts in oncology and artificial intelligence, including Dr. Fei Geng, recipient of 2023 Lung Ambition Award, thoracic oncologist Dr. Rosalyn Juergens, and AI scientist Dr. Zhiyong Zhang from Stanford University. Together, they are working to deliver new tools that put patients and families at the center of lung cancer care.

DETAILED BUDGET		**FROM:** **12/1/2025**	**TO:** **11/30/2026**
PERSONNEL		**SALARY REQUESTED**	
NAME	**ROLE ON PROJECT**	**Year 1**	**TOTALS**
Dr. Fei Geng	PI	$ 0 -	$ 0 -
Dr. Rosalyn Juergens	Co-PI	$ 0 -	$ 0 -
Dr. Zhiyong Zhang	Co-PI	$ 0 -	$ 0 -
Wenlong Wang	M.A.Sc. Student	$12,500	$12,500
PERSONNEL TOTAL		$12,500	**$12,500**
DATA ACCESS AND BIOMARKER ANALYSIS Funds are allocated for data access agreements or cost-recovery fees where required for controlled-access biobank datasets, while maximizing the use of freely available public resources.			**$5,000**
FACILITY USAGE GPU-accelerated cloud computing resources will be required for deep learning model training, validation, and interpretability analyses. Secure storage for clinical and biomarker data is also included.			**$5,000**
KNOWLEDGE TRANSLATION AND DISSEMINATION This will support preparation of manuscripts for submission to clinical oncology journals, conference abstract fees (e.g., WCLC 2026), and to be shared with Lung Cancer Canada and the patient community.			**$2,500**
TOTAL COSTS			**$25,000**

This pilot project has been carefully scoped to fit within the $25,000 budget while ensuring feasibility and high impact. The funds will be allocated as follows:

1. Personnel – $12,500

Wenlong Wang (M.A.Sc. student) will be supported to carry out data preprocessing, AI framework adaptation, and statistical analyses. Personnel costs represent the largest portion of the budget because specialized computational and analytic expertise is critical to the project's success. Investigators (Drs. Geng, Juergens, and Zhang) will contribute their time in-kind without requesting salary support.

2. Data Access and Biomarker Analysis – $5,000

Funds are allocated for data access agreements or cost-recovery fees where required for controlled-access biobank datasets, while maximizing the use of freely available public resources (e.g., TCGA, cBioPortal, ICGC). This line also covers limited targeted biomarker validation (such as ctDNA profiling for EGFR, KRAS, ALK, TP53 using digital PCR assays if required). These allocations ensure that the project remains feasible within the funding envelope.

3. Facility Usage (Computational Infrastructure) – $5,000

GPU-accelerated cloud computing resources are necessary for deep learning model training, validation, and interpretability analyses. Secure storage for clinical and biomarker data is included. Institutional infrastructure and in-kind computational resources will be leveraged to maximize efficiency.

4. Knowledge Translation and Dissemination – $2,500

This budget item supports the preparation of manuscripts for submission to peer-reviewed oncology journals, abstract submissions to national and international conferences (e.g., WCLC 2026), and lay summaries to be shared with Lung Cancer Canada and the patient community.

Total: $25,000 CAD

This budget emphasizes cost efficiency by leveraging open-access datasets, institutional infrastructure, and in-kind contributions from investigators. It is realistic, feasible, and aligned with the goals of the Give A Breath Research Award—to support a high-impact, proof-of-concept study that accelerates the translation of scientific discoveries into optimized patient care and improved outcomes.

McMaster University | Research Office for Administration, Development & Support

Gilmour Hall, Room 305
1280 Main Street West
Hamilton, ON, Canada, L8S 4L8
(905) 525-9140
https://roads.mcmaster.ca/

September 30, 2025

Dear Lung Cancer Canada:

McMaster enthusiastically supports the proposed research project by Dr. Fei Geng, associate professor at McMaster University. The project, titled “AI-Enhanced Liquid Biopsy for Monitoring Treatment Resistance in Advanced Lung Cancer” and submitted for consideration within the Give a Breath Research Award competition, is a compelling and innovative endeavor that aligns seamlessly with our institution's research objectives and goals.

Having thoroughly reviewed the details of the proposed research, we are confident in its feasibility within our institution. Our institution possesses the necessary infrastructure, resources, and expertise to facilitate the successful execution of this project. Furthermore, we acknowledge Dr. Fei Geng's expertise and dedication to their work. Their proven track record and commitment to excellence make us confident in their ability to carry out this research successfully. We anticipate that the outcomes of this research will not only enhance the academic reputation of our institution but also contribute meaningfully to the broader scientific community.

McMaster University intends to provide support for this project in the areas of grant fund administration, data management consultations, and institutional administrative support. We look forward to the positive impact Dr. Fei Geng’s research will have on our institution and the broader academic community.

Sincerely,
Sherrise Webb

Director, Research Office for Administration, Development and Support

Appendix: Supporting Manuscript

This Appendix provides an unpublished manuscript describing our development of a Discriminative Center-Loss Deep Learning framework for liquid biopsy–based lung cancer detection, offering preliminary evidence and methodological detail that support the feasibility and innovation of the proposed project.

Discriminative Center-Loss Deep Learning Enhances Liquid Biopsy-Based Early Detection of Lung Cancer

Wenlong Wang, Yijin Yang, Harmanpreet Multani, Fei Geng*

Faculty of Engineering, McMaster University, Hamilton, ON, L8S 0A3

*Corresponding Author, Fei Geng, gengf@mcmaster.ca

ABSTRACT

Liquid biopsy offers a non-invasive approach for early cancer detection by analyzing circulating tumor DNA (ctDNA) and cell-free DNA (cfDNA) in blood samples. However, the complexity and imbalance of clinical datasets challenge conventional models. Here, we implemented a Discriminative Center-Loss Deep Learning (DCLDL) framework to enhance classification performance in lung cancer detection from plasma-derived biomarkers. Using a curated dataset of 916 patients (812 healthy, 104 lung cancer), our model achieved 97% accuracy with an AUC of 0.994, outperforming a standard multilayer perceptron (MLP). Visualization of embedding spaces demonstrated improved intra-class compactness and inter-class separation, highlighting the interpretability advantages of DCLDL. These results underscore the promise of advanced AI frameworks in non-invasive diagnostics and pave the way for clinically relevant workflows in liquid biopsy.

INTRODUCTION

Lung cancer remains the leading cause of cancer-related mortality worldwide, with survival outcomes largely determined by the stage at which the disease is diagnosed. Early detection is critical, as patients diagnosed at localized stages experience markedly improved prognosis compared with those identified after disease progression. Conventional diagnostic approaches rely heavily on tissue biopsies and imaging modalities, which are invasive, costly, and often limited in their ability to capture tumor heterogeneity or detect disease at its earliest stages.

Liquid biopsy has emerged as a promising alternative, enabling minimally invasive detection and monitoring of circulating tumor DNA (ctDNA), cell-free DNA (cfDNA), and other blood-based biomarkers. Despite its potential, clinical adoption of liquid biopsy for lung cancer detection has been hindered by key challenges. These include the inherently high dimensionality and heterogeneity of biomarker datasets, low abundance of ctDNA in early-stage disease, and imbalances in patient cohorts that bias model performance. Traditional statistical and machine learning approaches have struggled to overcome these barriers, frequently resulting in suboptimal classification accuracy, poor reproducibility, and limited interpretability.

Recent advances in artificial intelligence (AI) offer new opportunities to address these limitations by extracting complex, high-order patterns from multi-omic data. In particular, deep learning models are well suited to capture nonlinear relationships among diverse biomarker inputs. However, conventional deep learning frameworks may still fail to achieve robust class separation when applied to imbalanced or noisy clinical datasets. To overcome these issues, we developed a Discriminative Center-Loss Deep Learning (DCLDL) framework that integrates a center-loss objective into neural network training. This approach enforces compact intra-class clustering and clear inter-class separability, thereby enhancing both predictive accuracy and interpretability.

Here, we applied the DCLDL framework to a curated liquid biopsy dataset for lung cancer detection. We demonstrate that this method achieves superior performance compared with standard multilayer perceptron models, attaining high accuracy and near-perfect area under the curve (AUC) values while producing interpretable feature embeddings. These findings highlight the promise of DCLDL as an AI-powered diagnostic tool and provide proof-of-concept evidence for its clinical potential in non-invasive early detection of lung cancer.

METHODS

Dataset and Preprocessing

Patient data were obtained from the publicly available liquid biopsy study by Cohen et al. (2018), which includes blood-derived molecular profiles used for cancer detection. For the present study, we focused specifically on lung cancer classification. The dataset comprised 916 samples, including 812 healthy controls and 104 patients diagnosed with lung cancer. Each sample contained multiple feature categories: plasma DNA concentration, mutation allele frequencies of circulating tumor DNA (ctDNA), over forty circulating protein biomarkers, and basic demographic variables such as age and sex.

To ensure data quality and consistency, preprocessing was conducted in several stages. First, categorical demographic features were encoded using one-hot encoding. Continuous features were standardized to zero mean and unit variance to prevent scale differences from biasing model training. Missing data points, which are common in clinical datasets, were addressed using k-nearest neighbors (KNN) imputation with k=5, allowing robust estimation of missing values based on the most similar samples. Following preprocessing, the dataset was randomly partitioned into training (70%), validation (15%), and test (15%) subsets, with stratification to maintain proportional representation of cancer and healthy cases across all splits.

Model Architecture and Training

We implemented a fully connected neural network augmented with a center-loss component, hereafter referred to as the Discriminative Center-Loss Deep Learning (DCLDL) model. The network architecture consisted of three hidden layers with rectified linear unit (ReLU) activations, batch normalization, and dropout regularization (p = 0.5) to reduce overfitting. The center-loss objective was integrated into the final feature embedding layer. This term penalizes the Euclidean distance between sample embeddings and their corresponding class centers, thereby enforcing tighter intra-class clustering and improved inter-class separation.

The final loss function was a weighted combination of cross-entropy loss and center loss:

$$L = L_{\text{softmax}} + \lambda L_{\text{center}}$$

where L_{softmax} is the standard cross-entropy loss, L_{center} is the center-loss term, and λ is a hyperparameter controlling their relative contributions.

Training was performed using the Adam optimizer for network weights (learning rate = 0.001) and stochastic gradient descent (SGD) for updating class centers (learning rate = 0.5). Early stopping was applied with a patience of 20 epochs based on validation loss to prevent overfitting. Training and evaluation were conducted using PyTorch 2.0 on an NVIDIA GPU workstation.

Baseline Model for Comparison

To establish a benchmark, we trained a multilayer perceptron (MLP) using the same architecture and hyperparameters but without the center-loss component. This baseline relied solely on cross-entropy loss for classification, enabling direct comparison of the effects of incorporating center loss. Both models were trained under identical preprocessing and data splits to ensure fairness of comparison.

Evaluation Metrics

Model performance was evaluated on the independent test set using accuracy, precision, recall, and F1-score for both cancer and healthy classes. Additionally, we computed the area under the receiver operating characteristic curve (AUC) to quantify overall discriminative performance. To better understand model robustness, we also examined confusion matrices and compared embedding spaces visualized using t-distributed stochastic neighbor embedding (t-SNE). Visualization was performed on the learned feature embeddings from the penultimate network layer, allowing assessment of intra-class compactness and inter-class separation.

Reproducibility and Validation

To ensure reproducibility, all experiments were repeated five times with different random seeds, and results are reported as the mean across runs. Hyperparameters, including the weighting factor λ\\lambdaλ, were tuned through grid search on the validation set. The final model configuration was selected based on the best balance of accuracy and AUC.

RESULTS

Training Dynamics and Model Convergence

The DCLDL model demonstrated stable convergence throughout the training and validation phases. As shown in Figure 1, both training and validation losses decreased steadily during the early epochs and plateaued at low values, indicating that the model avoided overfitting. Validation accuracy stabilized near 98–99% after approximately fifty epochs, confirming strong generalization to unseen data. The confusion matrix in Figure 1 further illustrates the classification performance, showing that the model correctly identified the vast majority of both healthy and cancer samples, with only a small number of cancer cases misclassified as healthy. This misclassification pattern likely reflects the underlying dataset imbalance, where the smaller cancer cohort constrained the model's ability to capture all disease-specific features.

Classification Performance

On the independent test set, the DCLDL framework achieved an overall accuracy of 97 percent, which represented an improvement over the baseline MLP model that achieved 94 percent. The receiver operating characteristic (ROC) analysis yielded an area under the curve (AUC) of 0.994, as illustrated in Figure 3, demonstrating near-perfect discriminative capacity. The model exhibited strong sensitivity and specificity. For cancer detection, the model attained a precision of 0.93 and a recall of 0.81, showing that most true cancer cases were identified correctly, while false positives were minimized. For healthy classification, the model achieved a precision of 0.98 and a recall of 0.99, reflecting exceptional reliability in distinguishing non-cancer samples. Although recall for the cancer class was slightly lower, this trade-off is consistent with the underrepresentation of cancer cases in the dataset. Notably, the high precision of the cancer predictions indicates that when the model classified a patient as cancer, the prediction was rarely incorrect, a clinically relevant property for diagnostic applications.

Comparative Analysis with MLP Baseline

Direct comparison with the baseline MLP highlights the advantages of the DCLDL approach. While the MLP achieved good performance with an AUC of 0.978 and an accuracy of 94 percent, it displayed greater variability in the classification of cancer cases and produced more false negatives. The embedding spaces produced by the two models further emphasize these differences. The t-SNE embeddings of the DCLDL model (Figure 2) show tight, well-separated clusters for healthy and cancer cases, while the MLP embeddings (Figure 4) appear dispersed with overlapping distributions. This demonstrates that the inclusion of the center-loss objective strengthened intra-class compactness and improved inter-class separability, yielding more robust and interpretable feature representations.

Feature Space Visualization and Interpretability

The two-dimensional visualization of the learned feature space underscores the interpretability benefits of the DCLDL framework. In Figure 2, cancer cases are clearly separated from the dense clusters of healthy cases, forming a distinct and interpretable pattern that aligns with biological expectations of disease-associated molecular signatures. In contrast, the MLP model (Figure 4) failed to achieve this separation, with substantial overlap between cancer and healthy samples. These visualizations provide an intuitive validation of the discriminative power of DCLDL, offering a transparent representation of the model's decision boundaries.

ROC Curve Analysis

The ROC curve for the DCLDL model, shown in Figure 3, provides further evidence of its discriminative capacity. The steep ascent of the curve and the near-maximal AUC value of 0.994 reflect strong sensitivity across a wide range of classification thresholds. Importantly, even at high-specificity thresholds, where false positives are minimized, the model maintained high sensitivity in detecting cancer cases. This balance between sensitivity and specificity underscores the clinical relevance of the approach, particularly in the context of screening applications where minimizing false positives while maintaining high cancer detection rates is essential.

DISCUSSION

The present study demonstrates that a Discriminative Center-Loss Deep Learning (DCLDL) framework can substantially improve classification performance in liquid biopsy–based lung cancer detection compared to a conventional multilayer perceptron. By integrating a center-loss objective, the DCLDL approach enforced intra-class compactness and inter-class separability, which translated into clearer feature clustering, higher predictive accuracy, and superior robustness. The combination of quantitative metrics and qualitative visualization provides convergent evidence that this framework is well suited for clinical applications in early cancer detection.

A key strength of the DCLDL model lies in its ability to mitigate challenges inherent to liquid biopsy datasets, including high dimensionality, noise, and class imbalance. While conventional models often achieve high training accuracy but generalize poorly to unseen data, the DCLDL model not only maintained strong performance on the validation and test sets but also produced well-structured feature embeddings. The t-SNE plots highlight that healthy and cancer cases form distinct distributions under the DCLDL model, which was not achieved by the baseline MLP. This improvement is particularly meaningful in the clinical setting, where interpretability and reproducibility are essential for building trust in AI-driven diagnostic systems.

The clinical implications of these findings are significant. A non-invasive, highly accurate, and interpretable diagnostic tool has the potential to complement or, in some cases, reduce reliance on invasive tissue biopsies and imaging modalities. The high precision observed for cancer classification suggests that the model rarely produces false cancer diagnoses, thereby minimizing unnecessary follow-up procedures. Although sensitivity for the cancer cohort was somewhat reduced due to class imbalance, the near-perfect AUC indicates that with better-balanced datasets, the framework could achieve even higher recall without compromising specificity. Such a balance is crucial in screening scenarios, where both false positives and false negatives carry serious clinical consequences.

Despite these strengths, several limitations must be acknowledged. The dataset included only 104 lung cancer cases compared to 812 healthy controls, creating a significant imbalance. While the center-loss approach helped mitigate this, the limited representation of cancer cases likely constrained the model's ability to capture the full heterogeneity of disease states. Additionally, the performance of deep learning models is sensitive to hyperparameter selection, and the λ weighting term in the center-loss function requires careful tuning to ensure reproducibility. These considerations highlight the need for validation across larger and more diverse cohorts that reflect real-world patient populations.

Conclusion

The DCLDL framework represents a significant step forward in AI-powered liquid biopsy diagnostics. By achieving high accuracy and interpretable cluster separation, it demonstrates potential as a clinically viable tool for early cancer detection. Future work should focus on larger, balanced datasets and interpretable workflows to accelerate clinical translation.

Figures

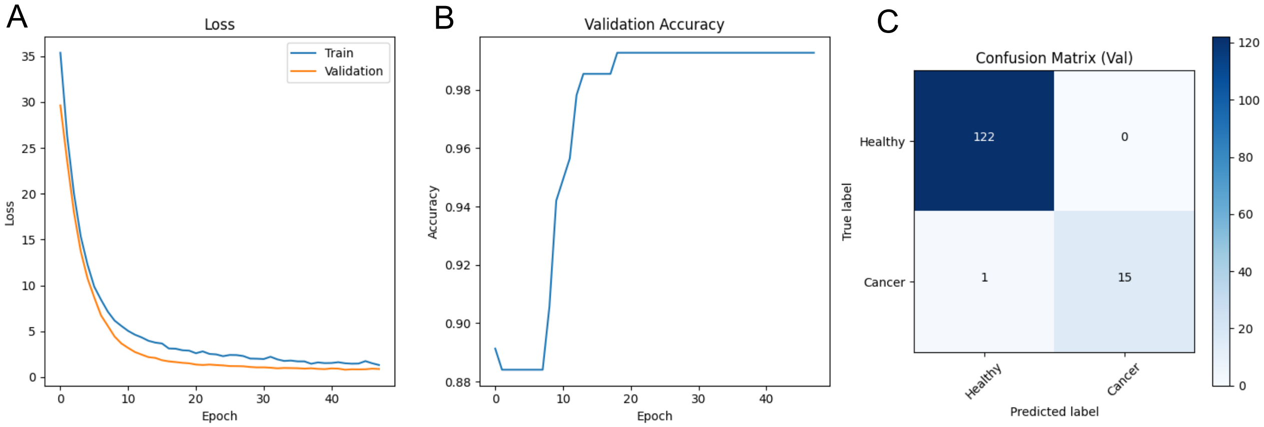

Figure 1. Training dynamics and validation performance of the DCLDL model. (A) Training and validation loss curves over 45 epochs, showing stable convergence and reduced overfitting. (B) Validation accuracy across epochs, plateauing near 99% after ~20 epochs, indicating strong generalization. (C) Confusion matrix of the validation set, demonstrating near-perfect classification with only one cancer case misclassified as healthy (accuracy = 98.8%).

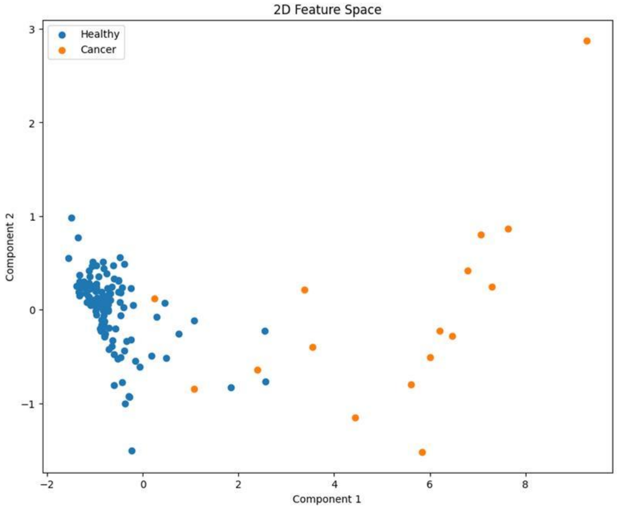

Figure 2. Two-dimensional feature space visualization of DCLDL embeddings. t-SNE projection of the learned feature representations shows clear separability between healthy (blue) and cancer (orange) samples, with compact intra-class clustering and distinct inter-class boundaries.

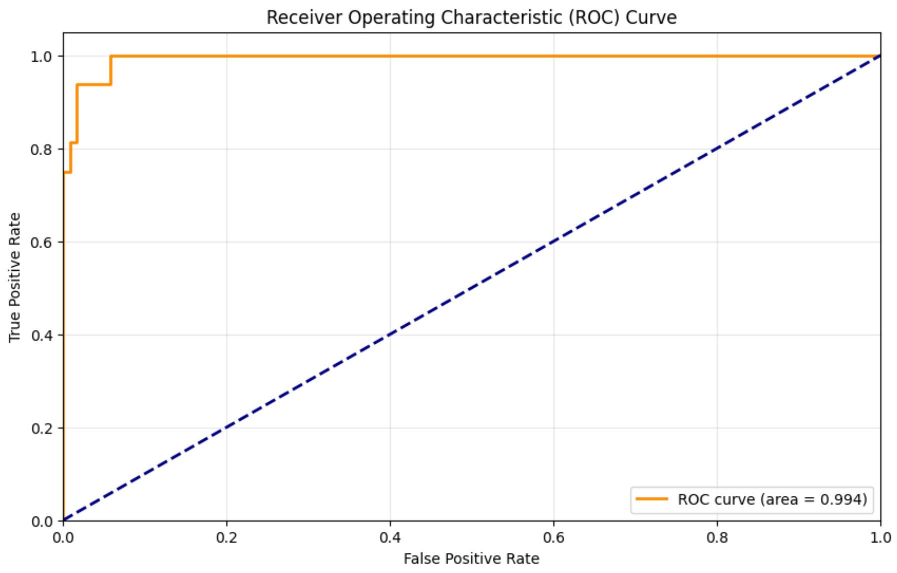

Figure 3. Receiver operating characteristic (ROC) curve of the DCLDL model. The ROC curve demonstrates excellent classification performance for distinguishing cancer from healthy samples, with an area under the curve (AUC) of 0.994, indicating near-perfect discriminative capacity.

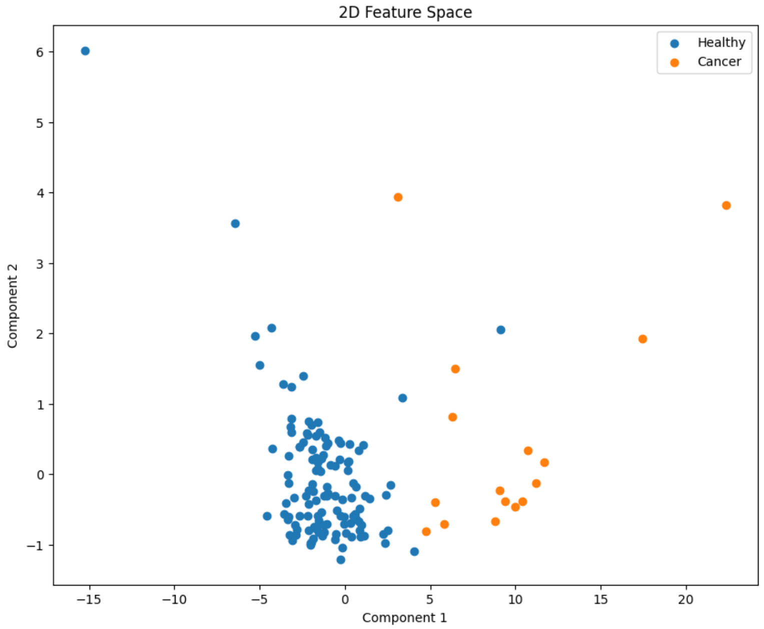

Figure 4. Two-dimensional feature space visualization of MLP embeddings. t-SNE projection of feature representations from the baseline multilayer perceptron (MLP) model shows dispersed clustering with partial overlap between healthy (blue) and cancer (orange) samples, indicating reduced class separability compared with the DCLDL framework.