Deep Learning Enhanced Vibrational Spectroscopy for Quantitative DNA Fragment Profiling

Rashad Fatayer*[a], Stephen John Sammut[b] , Ganapathy Senthil Murugan[a]
[a]Optoelectronics Research Centre, University of Southampton, Southampton, UK
[b]Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK

ABSTRACT

We demonstrate a deep learning approach for quantitative DNA fragment length analysis using vibrational spectroscopy. Controlled DNA mixtures spanning 50–300 bp were produced to establish distributions, enabling interpretable regression modelling. A convolutional neural network (CNN) with an attention module was trained to predict fragment proportions from spectral data. The model reliably reconstructed distributional differences, providing clear profiles that facilitate interpretation of biologically meaningful fragmentation patterns. To our knowledge, this is the first application of vibrational spectroscopy with deep learning for resolving DNA fragment length distributions, offering a rapid, label free, and non-destructive complement to existing molecular assays.

Keywords: Deep learning, machine learning, vibrational spectroscopy, ATR-FTIR spectroscopy, circulating tumor DNA, cell free DNA, cancer diagnostics, DNA

1. INTRODUCTION

Cell-free DNA (cfDNA) circulates in the bloodstream because of normal cellular turnover, with a small fraction in cancer patients originating from tumours [1]. This tumour derived fraction, often present at very low levels, has attracted major interest for its potential in minimally invasive liquid biopsies, enabling earlier detection and real-time disease monitoring [2]. Traditional methods such as digital polymerase chain reaction (dPCR) and next-generation sequencing (NGS) provide sensitive mutation profiling but remain limited by low cfDNA yields, high background from non-tumour cfDNA, and technical errors, particularly in early disease and minimal residual disease scenarios [3].

Fragment length analysis has emerged as a powerful complementary strategy. While cfDNA typically shows a modal size of ~166 bp, studies such as Florent et al. [4] have reported tumour derived DNA fragments to be significantly shorter (often 90–150 bp). Assessing these shifts in length distribution provides an additional line of evidence, helping to enrich tumour derived signals and improve detection sensitivity across a range of clinical contexts. Despite its promise, widespread adoption remains limited due to lack of standardization, increased cost, and added workflow complexity [5].

In this work, we introduce a novel approach for DNA fragment length analysis using attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy, with a convolutional neural network (CNN) enhanced by a squeeze-and excitation attention block. Unlike traditional vibrational spectroscopy applications that focus on binary classification of cancer versus healthy states, our framework applies regression-based quantification to directly resolve fragment length distributions. This strategy adds a new dimension to cfDNA analysis, with the potential to enhance liquid biopsy workflows through sensitive, label-free, and non-destructive DNA fragment profiling.

First Author, email: *rashad.fatayer@soton.ac.uk

Optics in Health Care and Biomedical Optics XV, edited by Qingming Luo, Xingde Li, Ying Gu, Dan Zhu,
Proc. of SPIE Vol. 13721, 1372113 · 2025 Published by SPIE
0277-786X · doi: 10.1117/12.3076499

2. METHODOLOGY

2.1 Materials

Synthetic DNA fragments of defined lengths (50, 100, 150, 200, and 300 base pairs) were obtained from ThermoFisher Scientific. Each fragment was supplied in Tris-EDTA (TE) buffer at a stock concentration of 0.5 μg/μL. Fragment length mixtures of varying proportions were prepared to reflect the heterogeneity of biologically derived cfDNA samples.

2.2 Spectroscopic measurements

Spectral acquisition was performed using an Agilent Cary 670 FTIR spectrometer equipped with a nine-bounce diamond-coated ZnSe MIRacle ATR accessory (PIKE Technologies, USA). For each measurement, 7 μL aliquots were deposited onto the ATR crystal and dried under ambient laboratory conditions prior to acquisition. Spectra were collected in the mid-infrared region (6000–700 cm^{-1}) at a resolution of 4 cm^{-1}, with 64 scans averaged per sample. An air background spectrum was acquired under identical conditions before each measurement to ensure spectral quality. To confirm reproducibility, each sample was measured in triplicate.

2.3 Data preprocessing

Raw ATR-FTIR spectra were preprocessed prior to analysis. The wavenumber range was restricted to the biomolecular fingerprint region (750–1800 cm^{-1}). Spectra were manually baseline corrected, denoised using a Savitzky–Golay filter with second-order derivative, and vector normalized to unit variance to reduce variability from film thickness and deposition.

2.4 Deep learning

A convolutional neural network (CNN) incorporating channel-wise attention (squeeze-and-excitation) for multiscale spectral feature learning was developed using Keras v3.7.0 with TensorFlow v2.17.0 as the backend, trained with the Adam optimizer and mean squared error (MSE) loss, and used early stopping to prevent overfitting. Figure 1 shows the CNN consists of two convolutional layers (64 and 96 filters with kernel sizes of 25 and 15, respectively), each followed by batch normalization and ReLU activation. These layers extract hierarchical spectral features, with the first convolutional block capturing broad baseline and peak structures and the second refining local spectral patterns. A squeeze–excitation (SE) block provides channel-wise attention, allowing the model to adaptively reweight informative features before global average pooling condenses them into a compact embedding. The dense head applies fully connected layers with ReLU activations and dropout for regularisation. The final layer outputs unconstrained regression values, which are subsequently rescaled to enforce mixture constraints, ensuring that predicted DNA fragment proportions remain physically consistent (non-negative and summing to 100%).

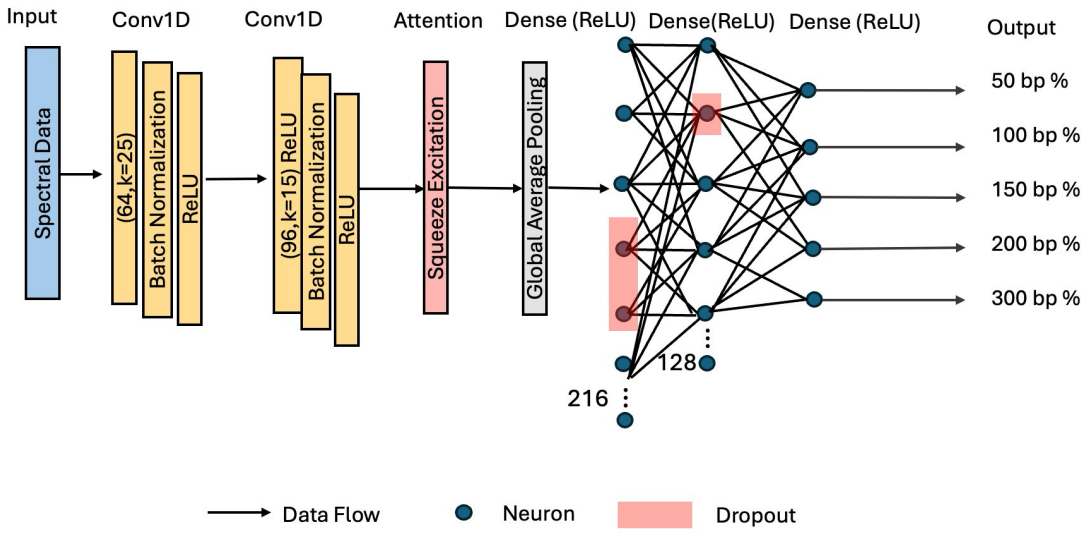

Figure 1: 1D CNN for DNA fragment length spectroscopic regression.

To expand the training set and improve model robustness, data augmentation was applied to each spectrum as shown in Figure 2, including small random baseline offsets, intensity scaling and additive noise, generating six augmented spectra per original sample. Hyperparameters were optimized using Optuna with Bayesian optimization, targeting minimization of validation loss, with searches executed in parallel on a high-performance computing (HPC) cluster equipped with NVIDIA L4 GPUs. Model performance was estimated using k-fold cross-validation with stratification by mixture composition to avoid data leakage and finally confirmed on a held-out test set. All model performance is reported using coefficient of determination (R^2), and root mean squared error (RMSE), calculated both per fragment and overall.

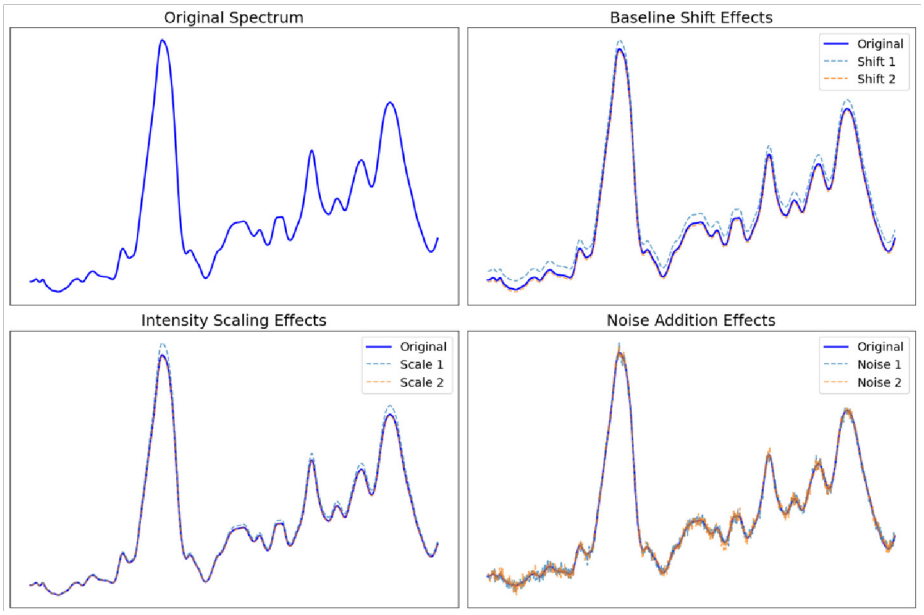

Figure 2: Visualization of data augmentation applied to DNA infrared spectra. Upper left: original spectrum; upper right: spectra with intensity scaling; lower left: spectra with baseline shifts; lower right: spectra with additive noise.

3. RESULTS

3.1 DNA fragment lengths

Infrared spectra obtained from DNA fragments of different lengths exhibit systematic variations that reflect changes in size.

3.2 Model training

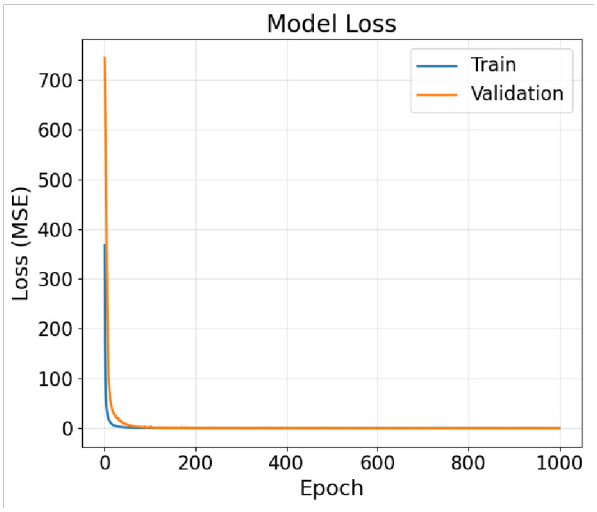

Figure 3: Training and validation loss during model optimization. The CNN was trained for 1000 epochs using the MSE as the loss function.

Model loss is shown in Figure 3 with a plot of training and validation mean squared error (MSE) over 1000 epochs. Both curves demonstrate a sharp decline during the initial ~80 epochs, after which early stopping was triggered to prevent overfitting. Beyond this point, the MSE values plateau, converging to consistently low levels with minimal separation between training and validation sets. The strong overlap of the two curves indicates effective generalization, with no evidence of overfitting or instability across the training process.

3.3 Mixture predictions

The regression analysis demonstrated accurate quantification of DNA mixtures, with strong predictive performance for shorter fragments (50–150 bp, $R^2 = 0.85–0.94$, RMSE = 4.9–7.5%) and reduced accuracy for longer fragments (200–300 bp, $R^2 = 0.65–0.76$, RMSE = 9.5–12.2%). When all fragment types were combined, the model maintained robust generalization ($R^2 = 0.83$, RMSE = 8.4%), confirming its suitability for predicting heterogeneous DNA mixtures.

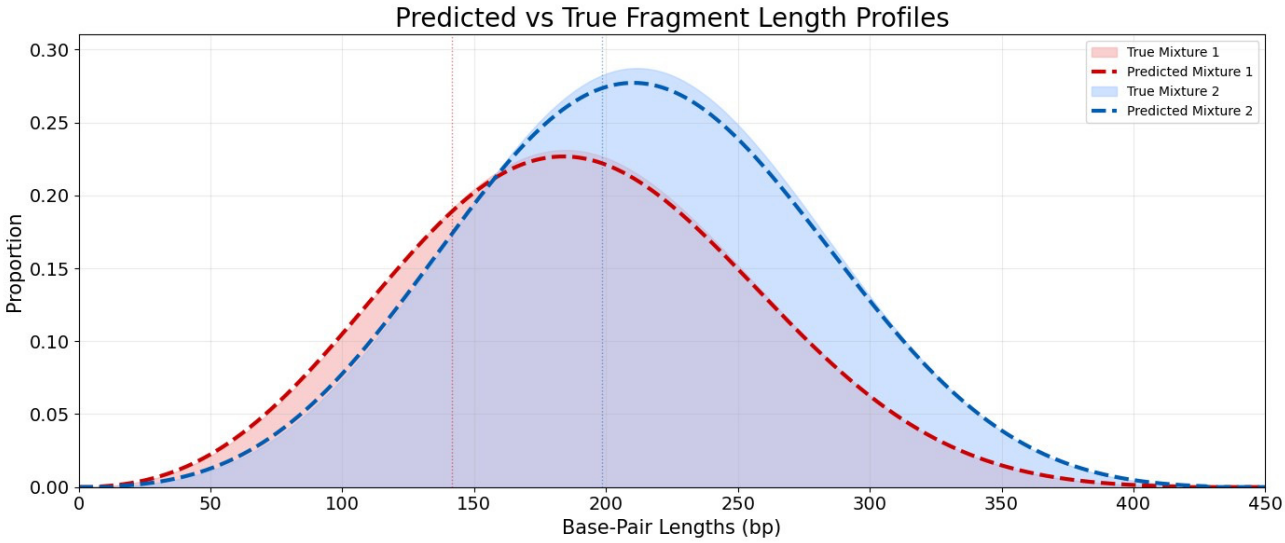

Figure 4: Comparison of predicted and true fragment length profiles for representative DNA mixtures. The solid curves represent the experimentally measured (true) fragment length profiles for two sample mixtures, while the dashed lines show the model predicted distributions generated using the CNN.

Figure 4 highlights the predicted fragment length distributions for two representative DNA mixtures. Sample 1 displays a dominant population of shorter fragments (~100 bp), whereas Sample 2 exhibits a broader distribution with a higher relative proportion of longer fragments (~250 bp). These differences arise from variations in the mixture compositions, reflecting changes in the relative abundance of fragment length populations. Visualizing DNA mixtures in this way enables clearer interpretation of compositional differences that may relate to underlying disease state.

4. CONCLUSION

This study demonstrates that vibrational spectroscopy combined with deep learning can quantify DNA fragment length distributions. The model achieved strong predictive performance, particularly for shorter fragments, reflecting the sensitivity of this approach to subtle spectral features associated with fragment size. Generating fragment length profiles from controlled DNA mixtures enables fully interpretable analysis and clear visualization of length dependent spectral variations. Collectively, these findings establish a foundation for extending the method to biological cfDNA samples, where differences in fragmentation profiles may hold diagnostic relevance.

ACKNOWLEDGEMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC grant EP/S03109X/1). RF thanks EPSRC DTP PhD studentship.

REFERENCES

[1] Sender, R., Noor, E., Milo, R. and Dor, Y., 2024. What fraction of cellular DNA turnover becomes cfDNA?. *Elife*, *12*, p.RP89321.

[2] Ma, L., Guo, H., Zhao, Y., Liu, Z., Wang, C., Bu, J., Sun, T. and Wei, J., 2024. Liquid biopsy in cancer: current status, challenges and future prospects. *Signal Transduction and Targeted Therapy*, *9*(1), p.336.

[3] Song, P., Wu, L.R., Yan, Y.H., Zhang, J.X., Chu, T., Kwong, L.N., Patel, A.A. and Zhang, D.Y., 2022. Limitations and opportunities of technologies for the analysis of cell-free DNA in cancer diagnostics. *Nature biomedical engineering*, *6*(3), pp.232-245.

[4] Mouliere, F., Chandrananda, D., Piskorz, A.M., Moore, E.K., Morris, J., Ahlborn, L.B., Mair, R., Goranova, T., Marass, F., Heider, K. and Wan, J.C., 2018. Enhanced detection of circulating tumor DNA by fragment size analysis. *Science translational medicine*, *10*(466), p.eaat4921.

[5] Guigal-Stephan, N., Lockhart, B., Moser, T. and Heitzer, E., 2025. A perspective review on the systematic implementation of ctDNA in phase I clinical trial drug development. *Journal of Experimental & Clinical Cancer Research*, *44*(1), p.79.