Notification No.: 591/2025
Date of award: 30-12-2025

Name of the Scholar: Syed Naseer Ahmad Shah

Student ID: 202008852

Name of the Supervisor: Prof. (Dr.) Rafat Parveen

Name of the Department: Department of Computer Science

Thesis Title: Deep Learning Approach to Enhance Lung Cancer Diagnosis Using Next Generation Sequencing

Findings

Lung cancer takes millions of lives around the world, mainly because it's often found late and is incredibly complex at the molecular level. This thesis uses the power of artificial intelligence, intense learning and bioinformatics to pick out critical genetic features to help doctors catch lung cancer earlier, sort tumours more accurately, and discover new genetic clues using big, modern sequencing datasets. The work introduces a smart combo of PCA (Principal Component Analysis) and MI (Mutual Information) to sift through genetic data, uses neural networks (CNNs) for precise predictions, and finally offers a new framework (called LUNGXAI) that not only predicts well but also explains itself, paving the way for more trustworthy AI-based cancer care.

Chapters 1 & 2: Introduction and Literature Review

These chapters open with global lung cancer statistics, underscoring its impact and the need for better diagnosis. The rapid progress in Next-Generation Sequencing (NGS) is highlighted, as it enables extensive genetic profiling but presents challenges due to massive data volume and complexity. The literature reviews the rise of AI in genomics, discussing the effectiveness of CNNs, GNNs, and modern variant callers in extracting value from biological data. The research objectives crystallise here, developing powerful feature extraction, boosting classification accuracy, and embedding model interpretability into lung cancer genomics.

Chapter 3: Biomarker Identification from RNA-Seq Data

This chapter shifts focus to discovering genes critical for early lung cancer detection. RNA-Seq datasets are analysed to identify upregulated (COL11A1, TOP2A, SULF1) and downregulated (PDK4, FOSB, MIR328) genes marking cancerous changes. Machine learning algorithms, Random

Forest, Lasso, and XGBoost, refine these gene sets for specificity. PPI network analysis then zeroes in on hub genes crucial to cancer progression, producing a targeted list of candidate biomarkers for future therapeutic intervention.

Chapter 4: Hybrid PCA-MI Feature Extraction Framework

Here, the thesis details the design of a novel hybrid feature extraction method that fuses Principal Component Analysis (PCA) and Mutual Information (MI). This approach solves the problem of high-dimensional gene data by retaining only features that are both discriminative and biologically relevant. Tested against TCGA and ICGC datasets, the PCA-MI framework consistently outperforms state-of-the-art techniques like Lasso, Autoencoders, and Random Forests for selection. Classifiers built on these features (notably CNNs) achieve 98% accuracy, and PPI validation confirms these genes are part of key cancer pathways, underscoring their clinical significance.

Chapter 5: LUNGXAI Interpretable Deep Learning Framework

This chapter introduces LUNGXAI, an advanced interpretable deep learning system tailored for lung cancer classification. The architecture combines robust CNN models with LIME-based interpretability. Built on GEO, TCGA, and ICGC datasets, LUNGXAI achieves 98.8% accuracy, outperforming traditional models. It also delivers explainable results by identifying top contributor genes (APP, DNMT3B, NFYC, TAF13), thus aligning model predictions with biological understanding, a critical requirement for clinical deployment of AI in medicine.

Chapter 6: Conclusion and Future Work

The final chapter synthesises the thesis's main achievements: the PCA-MI hybrid method, the accuracy of CNN classifiers on gene profiles, and the clinical relevance and transparency provided by LUNGXAI. Limitations are candidly addressed, specifically, the need for real-world clinical validation and broader dataset benchmarking. Future directions propose integrating multi-omics information, enabling pathway-level interpretations, and expanding explainable AI's role in tailoring oncology treatments for individual patients, pushing toward reliable clinical adoption.