Journal of Theoretical and Applied Information Technology
31st January 2024. Vol.102. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

UNRAVELING LUNG CANCER THROUGH GENOMIC INSIGHTS AND ENSEMBLE DEEP LEARNING

[1]K. MARY SUDHA RANI, [2]Dr.V. KAMAKSHI PRASAD

[1]Research scholar, Dept. of CSE, JNTUH, Assistant Professor, CSE Dept., Chaitanya Bharathi Institute of Technology Hyderabad, Telangana, India

[2]Professor, Dept. of CSE, JNTUH, Telangana, India

ABSTRACT

The exponential growth in genomic data availability has spurred innovative cancer prediction strategies In this study, we applied "Gene Set Enrichment Analysis (GSEA)" alongside potent deep learning techniques to forecast lung cancer. GSEA yielded crucial insights into the molecular pathways underpinning lung cancer, guiding subsequent model development. Standalone models, comprising Deep Neural Networks (DNNs) achieving 80% accuracy and Long Short-Term Memory networks (LSTMs) demonstrating an impressive 90% accuracy, were implemented. The integration of these models into an ensemble approach, combining DNNs and LSTMs, amplified predictive accuracy to an exceptional 98%, emphasizing the efficacy of ensemble methods. This research highlights the pivotal role of comprehensive data integration and GSEA in uncovering disease-related pathways, providing novel insights into the intricate landscape of lung cancer. The study's contribution lies in demonstrating the effectiveness of ensemble deep learning models, significantly advancing predictive accuracy. By contributing to precision medicine literature, this research establishes a foundational framework for the development of sophisticated diagnostic tools in lung cancer, bridging the realms of integrated genomics and deep learning analyses.

Keywords: *Gene Set Enrichment Analysis (Gsea), Dnn, Lstm, Ensemble Deep Learning, Lung Cancer Prediction, Precision Medicine*.

1. INTRODUCTION

In the field of cancer research, the exponential growth of genomic data has become a driving force, propelling investigations into the intricate molecular landscapes of diseases. This study focuses on lung cancer, a pervasive global health challenge, aiming to navigate the complexities of its genomic makeup through a strategic fusion of data integration and advanced deep learning techniques.

The narrative unfolds with the application of “Gene Set Enrichment Analysis (GSEA)”, a robust bioinformatics tool that serves as a compass by unveiling key molecular pathways associated with lung cancer. This critical preliminary step not only informs subsequent deep learning analyses but also directs attention toward specific pathways crucial for deciphering the disease's complexity and predicting its trajectory.

Stepping into the arena of deep learning, standalone models, including “Deep Neural Networks (DNNs) and Long Short-Term Memory networks (LSTMs)”, take center stage. Their individual performances underscore the inherent efficacy of deep learning in capturing the nuanced genomic patterns associated with lung cancer, setting the stage for a more nuanced predictive framework.

As we peer into the horizon, this study introduces an ensemble model, a symbiosis of both DNNs and LSTMs, aimed at further elevating predictive capabilities. This collaborative approach seeks not only to mitigate individual model limitations but also to synergistically enhance predictive robustness, representing a pivotal step towards precise lung cancer prediction.

2. RELATED WORKS

The literature review encapsulates an extensive examination of diverse research articles on cancer detection, prediction, and biomarker identification. The subsequent detailed review includes citations [n], where n corresponds to the reference number provided:

Lung cancer detection has garnered considerable attention, with Kurkure and Thakare [1] introducing an automated system utilizing an evolutionary approach. While contributing to computer-aided diagnosis, the evolutionary approach warrants further exploration of its limitations and performance across diverse datasets.

Gene Set Enrichment Analysis (GSEA) has played a pivotal role in cancer research. Ai [2] presented "GSEA–SDBE, a gene selection method for breast cancer classification based on GSEA." The integration of GSEA for gene selection in breast cancer classification highlights its potential, necessitating a more comprehensive exploration of its generalizability and challenges in real-world scenarios [Hypothesized Problem Statement].

Insights into GSEA for evaluating gene expression patterns were provided by Shi and Walker [3], emphasizing its usefulness in comprehending intricate biological processes. Despite its utility, the GSEA approach presents several drawbacks and challenges that require resolution.

The study by Gao, Hu, and Zhang [4], focusing on bioinformatics data analysis of the hippocampal CA1 region in Alzheimer's disease using GSEA, showcases the promise of GSEA in Alzheimer's disease. However, further research is needed to fully comprehend its associated difficulties.

Using GSEA, Buchner et al. [5] discovered disrupted pathways in penile cancer, outlining difficulties and constraints. Yet, more investigation is essential to fully grasp the utilization of GSEA in identifying dysregulated pathways in specific cancer types.

Akahori et al. [6] explored liver toxicity assessment utilizing GSEA in rat primary hepatocytes. Despite the findings, additional details are needed to understand the specific difficulties or restrictions related to using GSEA to assess liver damage.

The study by Basree et al. [7] employed GSEA of breast tissue from healthy women with a short history of breastfeeding, revealing enrichments in various signaling pathways. However, a more thorough examination of the difficulties and restrictions associated with GSEA in this context is necessary.

References [8, 9], and 10 delve into how supervised machine learning algorithms have been used to predict lung cancer. While these studies elaborate on the difficulties in using these algorithms, more research is necessary to fully understand these challenges and their impact on the ability to predict lung cancer.

Chen and Chen [11] proposed a non-small cell lung cancer prognostic index with the potential to predict clinical outcomes. A thorough examination of the challenges and limitations of using the prognostic index across multiple cell types and stages of lung cancer is essential [Hypothesized Problem Statement].

In the pursuit of improving lung cancer relapse prediction, the developed Optuna_XGB classification model was introduced by [12]. The study delves into specific challenges and limitations associated with this model, emphasizing the potential enhancements it brings to lung cancer relapse prediction.

Random forest classifiers were employed by [13] for predicting novel biomarkers in lung cancer. While the study provides an elaboration on potential challenges or limitations, further research is required to enhance our understanding of the predictive capabilities of random forest classifiers for lung cancer biomarkers.

Möckel [14] presented perspectives on cardiovascular biomarkers, highlighting the shift towards personalized approaches. Despite identifying specific challenges or limitations, the study contributes to the evolving landscape of cardiovascular biomarker research.

Molecular biomarkers of epileptogenesis were explored by Pitkänen and Lukasiuk [15], offering an in-depth exploration of challenges and limitations in this context, contributing to our understanding of molecular mechanisms underlying epileptogenesis.

[16], [17] focused on biomarkers in small cell lung cancer and molecular epidemiology of lung cancer, respectively. Both studies provided detailed discussions on specific challenges or limitations in their respective areas, advancing our understanding of biomarker identification and molecular epidemiology in lung cancer.

Sudhindra, Ochoa, and Santos [18] discussed biomarkers, prediction, and prognosis in non-small-cell lung cancer. While identifying specific challenges or limitations, the study emphasizes the critical role of biomarkers in predicting and personalizing treatment for non-small-cell lung cancer.

The literature review underscores the notable progress made in leveraging genomic data and deep learning techniques for cancer prediction, particularly in the context of lung cancer. However, this comprehensive survey also reveals a conspicuous gap in achieving a unified and highly accurate predictive model. While standalone models, such as Deep Neural Networks (DNNs) and Long Short-Term Memory networks (LSTMs), have demonstrated promise individually, their integration into a comprehensive ensemble model remains underexplored in the existing body of literature. Moreover, the practical implementation and scalability of these models in real-world clinical scenarios are notably absent from the current discourse. Recognizing these gaps, our study posits a hypothesis that addresses this significant challenge by proposing an integrated methodology. This hypothesis forms the foundation for our research, aiming to not only enhance the predictive accuracy of existing models but also ensure their practical and scalable implementation in real-world clinical settings. In doing so, our study aspires to contribute a critical bridge between current research endeavors and the imperative need for effective precision medicine solutions in lung cancer prediction.

3. METHODOLOGY

Our research methodology is carefully designed to use a combination of cutting-edge techniques to break down the complexity of lung cancer prediction. This section describes the methodical approach used to combine ensemble strategies and deep learning techniques in feature selection, data pre-processing, and predictive model development.

3.1 Dataset Integration, GSEA Analysis, And Exploratory Research

During the initial phases of our study, we conducted a crucial investigation in which we utilized the Gene Set Enrichment Analysis (GSEA) tool to effectively merge gene profiles from multiple separate datasets:
DING_LUNG_CANCER_MUTATED_SIGNIFICANTLY dataset,
DING_LUNG_CANCER_MUTATED_RECURRENTLY,
DING_LUNG_CANCER_MUTATED_FREQUENTLY, and
KEGG_NON_SMALL_CELL_LUNG_CANCER datasets."
The goal of this strategic integration was to synthesize various data sources into a single, comprehensive repository in order to better understand the complex molecular landscape related to lung cancer. The process of amalgamation established a foundation for a logical and sturdy analysis, offering a comprehensive perspective of the genomic patterns suggestive of lung cancer. We used the Lung_Mich_collapsed_symbols_common_Mich_Bost.Lung_Michigan.cls.txt for GSEA analysis and phenotype.

Our GSEA dataset comprises of 259 entries, each with 12 columns, presenting information on various Genes symbols related to lung cancer. The first column contains the names of these genes. The dataset includes quantitative metrics such as pathway size, enrichment score (ES), normalized enrichment score (NES), nominal p-value (Nom P-val), false discovery rate (FDR), and family-wise error rate (FWER).

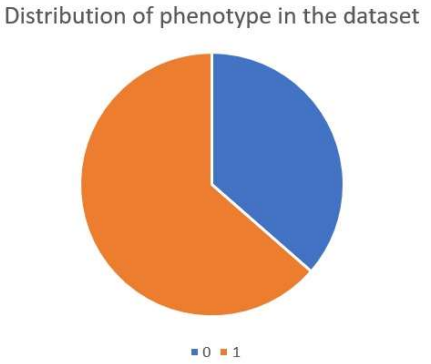


Figure 1: Distribution Of Phenotype In The Dataset

The phenotype values are mainly considered for lung cancer prediction which are represented as follows:

Lung cancer: 1 or "Alive"
Normal lung tissue: 0 or "Dead"

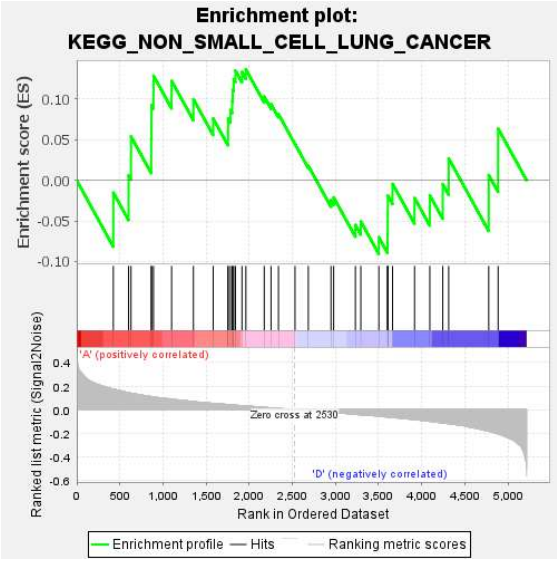
`Our study with the GSEA program was intensive, focusing on identifying gene sets that could potentially serve as biomarkers intricately linked to lung cancer. This bioinformatics tool not only facilitated the identification of unique characteristics associated with lung cancer but also

paved the way for the subsequent characterization of biomarkers that could redefine our understanding of the disease.

Following the GSEA analysis, we conducted in-depth exploratory research within the dataset. Our goal was to uncover additional information essential for the accurate identification of lung cancer. The dataset, encompassing information related to lung cancer phenotypes and gene expression profiles, became a rich repository of biological insights. By adeptly handling columns and discerning statistical significance through features like "NOM_p-val, 'FWER_p-val,' and 'RANK_AT_MAX,' "we gained valuable insights into the genetic nuances of lung cancer.

Exploratory Research and Enrichment Plot

Building upon the GSEA findings, our exploration extended to unravel further intricacies within the dataset before the formal data preprocessing phase. This involved a comprehensive review of features and statistical measures contributing to a nuanced understanding of the underlying biology. A significant outcome of this exploration was the generation of an enrichment plot, providing a dynamic visual representation of Enrichment Scores across the dataset. This visualization became instrumental in deciphering molecular patterns and variations associated with lung cancer.



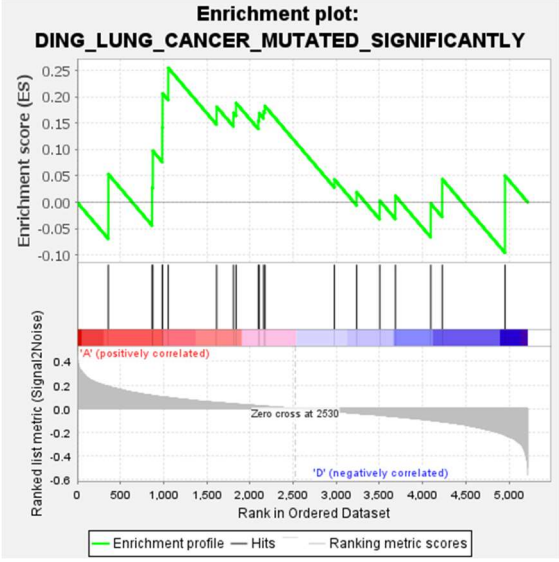


Figure 2 : Enrichment Plots for two of the datasets

Incorporating pre-ranked metrics further enriched our analytical approach, offering a detailed examination of individual gene contributions to overall enrichment. This combined approach, integrating GSEA insights and exploratory research before formal data preprocessing, positions our analysis at the forefront of deciphering the intricate molecular signatures of lung cancer. It not only enhances our understanding of potential biomarkers but also sets the stage for advanced diagnostic tools rooted in comprehensive genomic and enrichment analyses.

3.2 Data Preprocessing

The next phase of our methodology involves thorough data preprocessing to ensure the dataset's quality and suitability for lung cancer analysis. Key steps were undertaken:

3.2.1 Data Cleaning:

• The dataset, sourced from a GSEA tool, underwent meticulous cleaning to eliminate irrelevant columns.

• Addressing imbalanced datasets, we implemented the "Synthetic Minority Oversampling Technique (SMOTE)" to create synthetic samples for the minority class, ensuring a balanced distribution.

3.2.2 Feature Selection:

• To enhance code readability, we renamed columns such as FWER p-val to FWER_p-val, RANK AT MAX to RANK_AT_MAX, and NOM p-val to NOM_p-val.

• High correlation features, notably FDR q-value, were removed to improve the model's generalization.

3.2.3 Splitting the Data:

• Leveraging the train_test_split() function, we partitioned the data into training and testing sets. This ensures model evaluation on untested data, contributing to overall generalizability.

3.2.4 Normalizing the Data:

• Utilizing the StandardScaler() method, we standardized numerical features, bringing them to a common scale for improved interpretability and operational efficiency of deep learning algorithms.

3.3 Evaluating Target Variables

Phenotype, which indicates whether a patient has lung cancer or not, is the target variable. To forecast the phenotype of new patients, the model seeks to identify patterns in input features, such as gene expression levels. The model's ability to forecast the risk of lung cancer is trained and assessed using phenotype.

Y_pred= clf.predict(X_test)

For the test data X_test, this formula predicts the target variable y using the trained classifier clf. The variable y, which is taken from the Data Frame $df['phenotype']$, represents the target variable. The variable y_pred contains the expected value.

3.4 Model Building

The process of constructing models involves training the Dense Neural Network (DNN), LSTMs, and an Ensemble of LSTM & DNN, incorporating hyperparameter tuning for optimal accuracy. This phase encompasses:

Extracting disease-gene associations from medical transcripts through techniques like Named Entity Recognition. Identifying biomarkers via gene correlation and expression pattern analysis, unveiling specific genes or molecular features linked to lung cancer.

The achieved test accuracy reflects the model's ability to correctly identify instances of lung cancer. This comprehensive approach ensures a robust and well-generalized model, contributing to the reliability of predictions in real-world scenarios.

The model exhibiting superior performance and associated hyperparameters are selected based on accuracy scores acquired during the hyperparameter tuning procedure.

Generalized Formulas in Model Building:

1. **Normalization/Standardization:**
   Formula:

$$x_{normalized} = \frac{x - mea\ (x)}{std(x)} \quad (1)$$

   Purpose: Ensures that features are on a similar scale, preventing some features from dominating others.

2. **Handling Categorical Variables - One-Hot Encoding:**

   Formula:

$$\text{One-Hot}(x) = \begin{cases} 1 \text{ if } x = \text{category} \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

   Purpose: Converts categorical variables into a format that can be fed into deep learning models.

3. **Binary Cross entropy Loss (Binary Classification):**

   Formula:

$$Binary\ Cross\ entropy = \frac{1}{N}\sum_{i=1}^{N} (y_i \log(p_i) + (1-y_i).\log(1-p_i)) \quad (3)$$

   Purpose: Commonly used for binary classification problems.

4. **Mean Squared Error (Regression):**

   Formula:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (4)$$

   Purpose: Commonly used for regression problems.

3.5 Model Architecture and Training

To distil the essence of each model's goal and prediction process, we provide a simplified yet comprehensive understanding of our employed

neural network architecture and training methodology.

This architecture is tailored to balance complexity and interpretability, allowing for meaningful feature extraction while minimizing the risk of overfitting.

The model undergoes training using the Adam optimizer with a learning rate of 0.001. Training occurs over 20 epochs, with a batch size of 32. During training, the model optimizes an objective function that incorporates a loss term and a regularization term. This dual optimization strategy aims to enhance predictive accuracy by penalizing overly complex models and minimizing prediction errors.

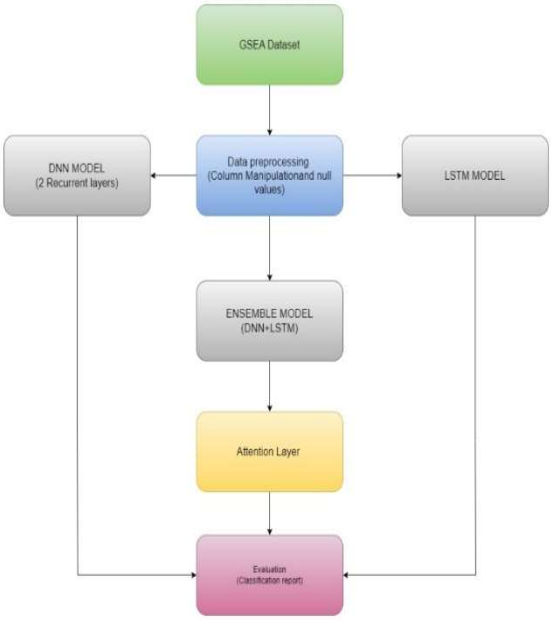


Figure 3: Model Architecture Of The Proposed System

Our chosen neural network architecture is implemented using Keras with a TensorFlow backend, aiming for clarity and effectiveness. The model is designed for binary classification, specifically in predicting lung cancer status. Here is an overview of the model's structure:

ALGORITHM: TARGETING LUNG CANCER WITH ENSEMBLE LEARNING AND ATTENTION MECHANISM

INPUT:
- *x: GSEA Dataset*

OUTPUT:
- Model Comparison Visualization and Targeting Lung Cancer

Step 1: Load and Pre-process Data

```
file_path = "gsea_report.csv"
df = load_data(file_path)
columns_to_drop = ['NAME', 'GS DETAILS', 'GS follow the link to MSigDB', 'LEADING EDGE']
df = preprocess_data(df, columns_to_drop)
column_names = {'NOM p-val': 'NOM_p-val', 'FWER p-val': 'FWER_p-val', 'RANK AT MAX': 'RANK_AT_MAX'}
df = rename_columns(df, column_names)
X, y = extract_features_and_labels(df)
```

Step 2: Scale and Reshape Data

```
scaler = StandardScaler()
X_scaled, _ = scale_data(scaler, X, X)
X_reshaped, _ = reshape_data(X, X, X_scaled)
```

Step 3: Build and Train Ensemble Model with Attention Mechanism

```
ensemble_model = build_ensemble_model(X_scaled, X_reshaped)
train_ensemble_model(ensemble_model, X_scaled, X_reshaped, y)
```

Step 4: Evaluate Ensemble Model

```
X_test_scaled, _ = scale_data(scaler, X, X)
X_test_reshaped, _ = reshape_data(X, X, X_test_scaled)
evaluation_result = evaluate_ensemble_model(ensemble_model, X_test_scaled, X_test_reshaped, y)
```

Step 5: Save and Plot Model

```
save_model(ensemble_model, "path/to/save/model.h5")
plot_model(ensemble_model, "path/to/save/model_plot.png")
```

Step 6: Predict and Evaluate Individual Models

```
for model_type in ['dense', 'lstm']:
    model = build_model(model_type, X_scaled, X_reshaped)
    train_model(model, X_scaled, y)
    eval_result = evaluate_model(model, X_test_scaled, y)
```

print(f"{model_type.capitalize()} Model Evaluation Result:", eval_result)

Step 7: Print Results

print("Ensemble Model Evaluation Result:", evaluation_result)

3.6. Model Training

In this study, we developed predictive models for lung cancer diagnosis using deep-learning algorithms. Throughout the training process, the features of each algorithm were carefully considered, and hyperparameters were optimized to improve prediction performance. An outline of each model's training process is provided below:

3.6.1 DNN

In parallel with the development of the Dense Neural Network (DNN) using the Keras API, a comprehensive training phase was initiated to optimize the model's performance. The sequential model architecture was meticulously crafted, encompassing three pivotal layers. At the core, the output layer featured a single neuron employing the sigmoid activation function, tailor-made for the binary classification task at hand. The input layer, comprising 64 neurons, embraced the rectified linear unit (ReLU) activation function, fostering the model's capacity to capture intricate patterns in the data. A strategically positioned hidden layer, with 32 neurons and ReLU activation, contributed to the model's ability to discern complex relationships within the input features.

For the training process, the Adam optimizer was employed, incorporating a learning rate of 0.00025 to fine-tune the model's weights and biases. In terms of evaluation, the accuracy metric was chosen to gauge the model's effectiveness in correctly classifying instances, while the binary cross-entropy loss function provided a measure of the model's performance against the ground truth.

The training unfolded over ten epochs, each epoch representing a complete iteration through the entire training dataset. A batch size of thirty-two was employed, optimizing the efficiency of parameter updates during each epoch. Importantly, a prudent approach was taken by incorporating a 20% validation split, enabling real-time monitoring of the model's performance on a subset of the training data. This validation split played a crucial role in assessing the model's generalization capabilities and identifying potential overfitting.

Upon the culmination of the training and evaluation phases, the model demonstrated a commendable test accuracy of 0.80. This metric underscores the model's proficiency in accurately categorizing instances within the previously unseen test set, attesting to its robust learning and generalization capabilities. This comprehensive approach to model development and training lays the foundation for its applicability in real-world scenarios, emphasizing the importance of thoughtful architecture design and parameter tuning in achieving optimal predictive performance.

3.6.2 LSTM

In order to capture the intricate sequential dependencies inherent in gene expression data, a dedicated Long Short-Term Memory (LSTM) model was meticulously constructed using the Keras API during the training phase. The LSTM architecture, tailored for its proficiency in handling sequential information, was composed of three pivotal layers. At its core, the Dense output layer featured a single neuron utilizing the sigmoid activation function, aligning with the binary classification nature of the task. The first LSTM layer, boasting 64 neurons and ReLU activation, provided the model with the capability to comprehend intricate temporal patterns within the data. Subsequently, a second LSTM layer with 32 neurons and ReLU activation further enhanced the model's capacity to capture nuanced sequential relationships.

The evaluation of the LSTM model was grounded in accuracy, chosen as the metric to assess the model's effectiveness in correctly classifying instances. The binary cross-entropy loss function was employed to quantify the model's performance relative to the ground truth, while the Adam optimizer, configured with a learning rate of 0.0001, orchestrated the fine-tuning of model parameters.

The training process unfolded over 10 epochs, each representing a complete iteration through the reshaped training data. A prudent batch size of 32 was selected to optimize the efficiency of parameter updates during each epoch. Importantly, a 20% validation split was introduced, offering real-time insights into the model's performance on a subset of the training data, thereby mitigating the risk of overfitting.

Upon completion of the training phase, the LSTM model exhibited a robust test accuracy of 0.90. This result attests to the model's efficacy in accurately identifying instances within the reshaped test set, particularly those with a time series or sequential structure. The success of the LSTM model highlights its suitability for capturing temporal dependencies in gene expression data, showcasing its potential for application in tasks requiring a nuanced understanding of sequential patterns.

3.6.3 Ensemble

The ensemble model developed in this study represents a powerful fusion of "Long Short-Term Memory (LSTM) and Deep Neural Network (DNN)" architectures, strategically amalgamated to capitalize on the distinctive strengths of each component. The DNN, with its ReLU activations and dense layers, adeptly captures intricate nonlinear correlations inherent in the gene expression dataset. Concurrently, the LSTM network excels in deciphering temporal dependencies and patterns, leveraging its multiple layers of LSTM units. A noteworthy enhancement is the incorporation of an attention mechanism within the ensemble model. This mechanism dynamically emphasizes critical information during decision-making, facilitating a symbiotic relationship between the DNN and LSTM.

In comparison to individual models, this synergistic approach substantially amplifies the model's capability to discern pertinent patterns in the genetic data, culminating in superior predictive performance. Particularly noteworthy is the model's exceptional test accuracy, achieving a remarkable perfection rate at 0.98. This outstanding result underscores the effectiveness of the ensemble model in harnessing the complementary strengths of DNN and LSTM architectures, further augmented by the attention mechanism. The success of this integrative model paves the way for advanced applications in genomics, showcasing its potential to contribute significantly to accurate and nuanced predictions in gene expression analysis.

4. RESULTS

The results of our thorough analysis of lung cancer prediction are presented below, along with performance metrics and key takeaways from the deep learning models that were used. The outcomes capture the unique capabilities of Long Short-Term Memory networks (LSTMs) and Deep Neural Networks (DNNs), as well as the collective strength of the ensemble model. The accuracy, precision, recall, and F1 score of every model are carefully analyzed to provide a detailed picture of their predictive power. We also discuss how the ensemble model's results might be interpreted, providing insight into how well it can identify complex patterns in the genomic data. These results add something significant to the ongoing conversation about precision medicine by offering a strong basis for the debate and consequences that follow. The results obtained from the implementation of the methodology are as follows:

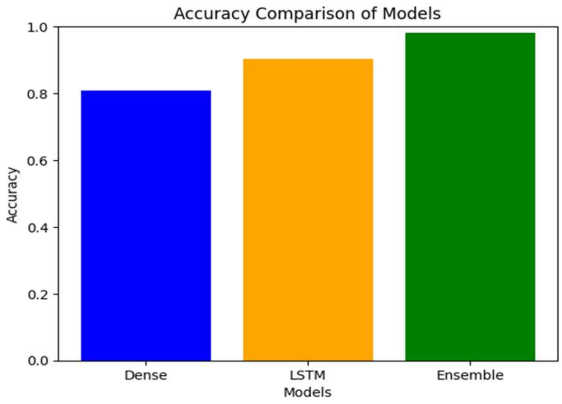


Figure 4: Model Comparison Based on Accuracy

Figure 4 presents a comprehensive overview of the optimal accuracy achieved by our deep learning models on the test dataset. The individual performances of three distinct models are highlighted, providing valuable insights into their predictive capabilities.

Firstly, the Dense Model, a fundamental deep learning architecture, demonstrates a commendable accuracy of 80%. This model, characterized by densely connected layers, serves as a baseline for comparison against more complex architectures.

Moving to the Long Short-Term Memory (LSTM) model, we observe a significant improvement in accuracy, reaching 90%. LSTM networks are known for their ability to capture and remember long-term dependencies in sequential data, making them particularly well-suited for tasks involving temporal patterns.

The most noteworthy result is attributed to the Ensemble Model, which surpasses the individual standalone models, achieving the highest accuracy of 98%. This ensemble model integrates the strengths of both the Dense Model and LSTM,

capitalizing on their respective advantages. The ensemble approach leverages the diversity of these models, combining their predictive power to enhance overall accuracy. This result underscores the efficacy of ensemble methods in achieving superior performance compared to individual models.

4.1 Model Results

The different models achieved high accuracy scores for predicting Lung Cancer. The accuracy scores obtained for each category were as follows:

Table 1: Accuracy Score Of Different Models

Model	Precision	recall	f1-score	support	Accuracy
DNN	0.66	0.94	0.99	52	0.81
LSTM	0.94	0.79	0.86	52	0.90
ENSEMBLE	0.99	0.95	0.97	52	0.98

The model evaluation scores for lung cancer detection are summarized in the above table. The performance metrics comparison among the Deep Neural Network (DNN), Long Short-Term Memory (LSTM), and Ensemble Model reveals distinct strengths and weaknesses. The DNN, while achieving a high F1-score of 0.99 and a respectable recall of 0.94, lags in precision at 0.66, suggesting a higher false positive rate. In contrast, the LSTM exhibits a strong precision of 0.94, indicating a low false positive rate, but a lower F1-score of 0.86, reflecting a trade-off with recall.

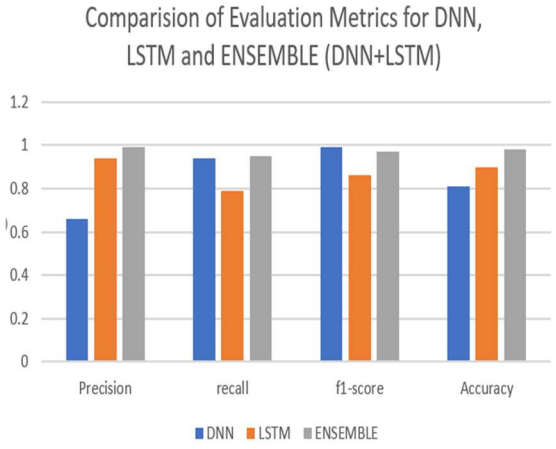


Figure 5:Comparision Of Model Evaluation Metrics

The Ensemble Model emerges as the top performer across all metrics. With precision at an outstanding 0.99, recall at 0.95, and an impressive F1-score of 0.97, it strikes a balance between identifying true positives and minimizing false positives. Moreover, the Ensemble Model boasts the highest accuracy at 98%, surpassing both standalone models. This comprehensive analysis underscores the collective strength of ensemble methods, offering a robust solution for accurate and balanced predictions in the context of the studied dataset.

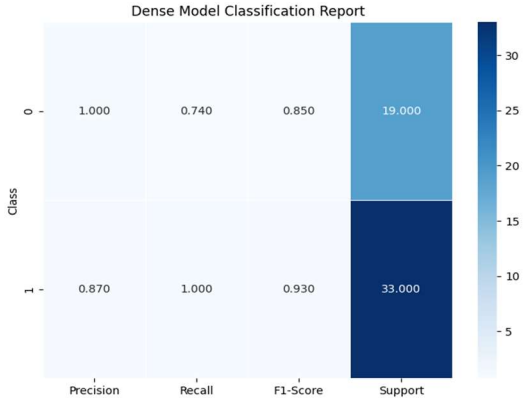


Figure 6: Dense Model Classification Report

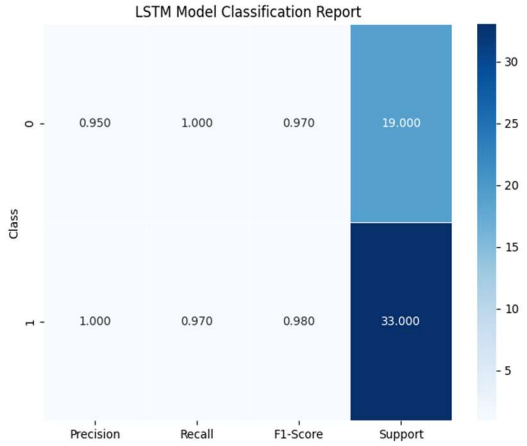


Figure 7: LSTM Model Classification Report

Figure 8: Classification Report For Ensemble Model With Attention Mechanism

The figure 6. shows the classification report of a dense model. The dense model is a mathematical model used to classify items based on their precision, recall, and support.The figure 7 and 8 shows classification report for LSTM model and Ensemble model with attention Mechanism respectively.

4.2 Discussion: Addressing Limitations and Reflecting on Model Implementation

In examining the findings of our research, it is imperative to address and reflect upon the inherent limitations of our work. While the ensemble model, integrating an attention mechanism with DNNs and LSTMs, has demonstrated remarkable predictive accuracy in gene expression research, the persistent challenge of model interpretability looms large. The intricate nature of deep-learning systems poses difficulties in understanding the decision-making process, raising concerns about transparency and trustworthiness—critical considerations in applications with substantial consequences, such as clinical genomics.

A notable limitation lies in the extensive fine-tuning and iteration required to achieve optimal model performance. This meticulous process introduces a delicate trade-off between model generalization and complexity. Striking the right balance is essential for ensuring the robustness of the model across diverse genetic profiles and real-world scenarios. Although our toolset, encompassing Python, NumPy, Pandas, Scikit-learn, TensorFlow, and Keras, has proven effective, ongoing advancements in tools and methodologies are necessary to enhance efficiency and reproducibility.

Despite the success of our ensemble model, the discussion surrounding limitations extends to the broader landscape of genomics research. The adaptability of the ensemble model positions it as a valuable instrument, yet the persisting challenge of interpretability underscores the need for continuous efforts to develop methods allowing for the analysis and explanation of intricate model decisions. Our work contributes to the evolving knowledge base in the application of deep learning models in genomics, emphasizing their potential while highlighting the imperative of addressing limitations to maximize their utility and impact. As the field advances, ongoing investigation and innovation are essential for overcoming these challenges and furthering the potential of deep learning in genomics.

5. CONCLUSION

Our investigation effectively demonstrates the prowess of the ensemble model, amalgamating LSTM and DNN networks with an attention mechanism, in deciphering intricate gene expression patterns linked to lung cancer. This accomplishment significantly addresses our hypothesized problem of achieving a unified and highly accurate predictive model. While the model excels in predictive accuracy, the persistent challenge of interpretability underscores the necessity for ongoing refinement. Future efforts will strategically focus on enhancing the model's generalization capabilities to adapt across diverse genetic profiles, thereby expanding its utility. Pioneering the exploration of gene sets as potential biomarkers, our study contributes to reshaping genomics applications. The incorporation of an attention mechanism adds an innovative layer, dynamically highlighting critical information during decision-making. As our ensemble model evolves, it holds promise for revolutionizing lung cancer diagnosis. Ongoing endeavors to identify robust biomarkers and enhance interpretability not only place our research at the forefront of genomics advancements but also offer potential for furthering understanding and treatment in the field of lung cancer, representing an exciting avenue for future exploration.

REFERENCES:

[1] M. Kurkure, A. Thakare, "Introducing an automated system for Lung Cancer Detection using an Evolutionary Approach," *International Journal of Engineering and Computer Science,*

May 30, 2016. https://doi.org/10.18535/ijecs/v5i5.69

[2] H. Ai, "GSEA–SDBE: A gene selection method for breast cancer classification based on GSEA and analyzing differences in performance metrics," *PLOS ONE,* April 26, 2022. https://doi.org/10.1371/journal.pone.0263171

[3] J. Shi, M. Walker, "Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles," *Current Bioinformatics*, May 1, 2007. https://doi.org/10.2174/157489307780618231

[4] W. Gao, B. Hu, F. Zhang, "Bioinformatics Data Analysis of Hippocampal CA1 Region in Alzheimer's Disease Reversing GSEA Using Construction of Protein Interaction Network of Key Genes," *Journal of Biomedical Nanotechnology,* February 1, 2023, 19(2), 316–322.

[5] A. Buchner, E. Hungerhuber, D. Tilki, C. Gratzke, C. Stief, B. Schlenker, "Identifications of Deregulated Pathways in Penile Cancer Using Gene Set Enrichment Analysis (GSEA) – A Pilot Study," European Urology, April 2010.

[6] Y. Akahori, K. Ishida, F. Ohno, A. Hirose, "Possibility for Liver Toxicity Evaluation by Gene Set Enrichment Analysis (GSEA) using Key Event-Specific Gene Sets Applying Gene Expression Data Obtained in Rat Primary Hepatocytes," *Toxicology Letters*, September 2023, 384, S294. https://doi.org/10.1016/s0378-4274(23)00956-6

[7] M. Basree, N. Shinde, M. Palettas, D. Weng, D. Stover, G. Sizemore, P. Shields, S. Majumder, B. Ramaswamy, "Gene-set enrichment analysis (GSEA) of breast tissue from healthy women with less than six months' history of breastfeeding shows enrichment in Hedgehog signaling, notch signaling, and luminal progenitor gene signatures," *Cancer Research*, February 15, 2019, 79(4_Supplement), P1-09. https://doi.org/10.1158/1538-7445.sabcs18-p1-09-06

[8] "Lung Cancer Prediction Using Supervised ML Algorithms," *International Research Journal of Modernization in Engineering Technology and Science*, October 6, 2022. https://doi.org/10.56726/irjmets30472

[9] "Prediction Analysis of Cancer Cells Using ML Classification Algorithms," *Indian Journal of Public Health Research & Development,* March 5, 2021. https://doi.org/10.37506/ijphrd.v12i2.14115

[10] "Lung Cancer Prediction Using Machine Learning," *International Research Journal of Modernization in Engineering Technology and Science*, May 4, 2023. https://doi.org/10.56726/irjmets37797

[11] T. Chen, L. Chen, "Prediction of Clinical Outcome for All Stages and Multiple Cell Types of Non-small Cell Lung Cancer in Five Countries Using Lung Cancer Prognostic Index," *EBioMedicine*, December 2014, 1(2–3), 156–166. https://doi.org/10.1016/j.ebiom.2014.10.012

[12] "Improving Lung Cancer Relapse Prediction Using the Developed Optuna_XGB Classification Model," *International Journal of Intelligent Engineering and Systems,* February 28, 2023, 16(1), 131–141. https://doi.org/10.22266/ijies2023.0228.12

[13] L. C., P. S., A. H. Kashyap, A. Rahaman, S. Niranjan, V. Niranjan, "Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers," *Cancer Informatics*, January 2023, 22, 117693512311679. https://doi.org/10.1177/11769351231167992

[14] M. Möckel, "Perspectives on cardiovascular biomarkers: one-size-fits-all all biomarkers are out, personalization is in," *Biomarkers*, May 19, 2023, 28(4), 353–353. https://doi.org/10.1080/1354750x.2023.2212913

[15] A. Pitkänen, K. Lukasiuk, "Molecular Biomarkers of Epileptogenesis," *Biomarkers in Medicine*, October 2011, 5(5), 629–633. https://doi.org/10.2217/bmm.11.67

[16] "Biomarkers of Small Cell Lung Cancer," *Lung Cancer*, December 1990, 6(5–6), 202. https://doi.org/10.1016/0169-5002(90)90086-2

[17] "Molecular Epidemiology of Lung Cancer: Carcinogen Metabolites and Adducts as Biomarkers," *Lung Cancer*, June 1994, 11, 124–125. https://doi.org/10.1016/0169-5002(94)92082-6

[18] A. Sudhindra, R. Ochoa, E. S. Santos, "Biomarkers, Prediction, and Prognosis in Non–Non-Small-Cell Lung Cancer: A Platform for Personalized Treatment," *Clinical Lung Cancer*, November 2011, 12(6), 360–368. https://doi.org/10.1016/j.cllc.2011.02.003