Building and training a neural network model to classify molecular subtypes of head and neck cancer based on gene expression.

Project submitted to the
SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering School of Engineering and Sciences

Submitted by

MAHESH TALLURI (AP21110010744)

CHARAN SAI.V (AP21110010662)

AKSHITH REDDY.G (AP21110011225)

ROHITHA (AP21110010710)

Under the Guid DR. ANIRBAN BHAR

SRM University–AP Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240 Nov, 2023

Table of Contents

Certificate

This is to certify that the work present in this Project entitled "**Neural Network Classification of Head and Neck Cancer Molecular Subtypes Using Gene Expression Profiling**" has been carried out by **[V.CHARAN SAI,L.ROHITHA,T.MAHESH,G.AKSHITH]** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

Supervisor

(signature)

Dr.Anirban Bhar

Assistant Professor

Dep.of CSE

SRM University, AP

ACKNOWLEDGEMENTS

We express our gratitude to Dr.Anirban Bhar for his individual guidance expertise, and unwavering support throughout the research process. His insightful inputs, encouragement, and commitment to excellence have significantly enriched this work on "Building and training a neural network model to classify molecular subtypes of head and neck cancer based on gene expression levels". Dr. Anirban Bhar's mentorship has been instrumental in shaping the direction and quality of this research.

Abstract

This report explores the application of neural network algorithms in classifying molecular subtypes of head and neck cancer based on gene expression profiling. The study aims to enhance our understanding of the heterogeneity within head and neck cancers, facilitating more personalized and effective treatment strategies.

Our model contributes to precision medicine by facilitating targeted therapeutic interventions, offering potential benefits in early detection and treatment planning. The research aligns with advancements in understanding the molecular basis of head and neck cancer, providing a tool for clinicians to make informed decisions. The integration of this model into diagnostic workflows promises to enhance efficiency and accuracy, fostering improved patient outcomes in the realm of head and neck cancer diagnostics and treatment strategies.

This study introduces a neural network model for classifying molecular subtypes of head and neck cancer based on gene expression levels. The model, trained on a comprehensive dataset, aims to enhance precision medicine, early detection, and clinical decision-making. The research contributes to the understanding of the molecular basis of head and neck cancer and offers a valuable tool for clinicians, potentially improving patient outcomes.

ABBREVIATIONS

NN: Neural Network
RNA-seq: RNA sequencing (a technique for gene expression analysis)
ReLU: Rectified Linear Unit (common activation function)
DNN: Deep Neural Network
CNN: Convolutional Neural Network (if using spatial relationships in gene expression)

- **Introduction**

Head and neck cancer comprises a diverse spectrum of malignancies, each characterized by unique molecular features that influence disease progression and treatment outcomes. In recent years, the advent of high-throughput gene expression profiling has revolutionized our ability to dissect the intricate molecular landscape of cancers. This report delves into the application of advanced neural network algorithms to classify the molecular subtypes of head and neck cancer based on comprehensive gene expression profiles. By harnessing the power of artificial intelligence, this study aims to unravel the complexities of head and neck cancer heterogeneity, paving the way for more refined diagnostic and therapeutic strategies tailored to individual patients.

Gene expression profiling enables the simultaneous assessment of thousands of genes, providing a comprehensive molecular snapshot of the tumor. This wealth of information can be leveraged to identify distinct molecular subtypes within HNC, each with unique biological characteristics and clinical behaviors. By employing advanced machine learning techniques, particularly neural networks, we aim to develop a robust and accurate model for classifying molecular subtypes of head and neck cancer based on gene expression data.

Objective:

The objective is to develop a neural network model that analyzes gene expression data to classify molecular subtypes of head and neck cancer. Through training, the model learns patterns in gene expression levels associated with different subtypes, enabling accurate classification and aiding in understanding the molecular basis of the cancer's heterogeneity.

The primary objective of this study is to build and train a neural network model capable of accurately classifying molecular subtypes of head and neck cancer based on gene expression profiles. By leveraging the power of artificial intelligence, we aim to enhance the precision of subtype classification, potentially uncovering novel molecular signatures that have diagnostic, prognostic, and therapeutic implications.

List of Equations

Cost Function:
J(W,b)=−m1∑i=1m∑c=1Cyc(i)log(ac(i)), where m is the number of samples, C is the number of classes, yc(i) is the ground truth label, and ac(i) is the predicted probability.

Methodology

1. Data Collection:

Dataset Description: Provide details about the dataset used, including the source, size, and any preprocessing steps applied.
Gene Expression Data: Specify the gene expression data used, highlighting the relevant genes and their expression levels.

2. Data Preprocessing:

Normalization: Normalize the gene expression data to ensure consistent and comparable values across samples.
Feature Selection: Identify and select relevant features, possibly using techniques such as variance thresholding or dimensionality reduction.

3. Neural Network Architecture:

Model Selection: Choose a neural network architecture suitable for the task. Consider architectures such as feedforward neural networks, convolutional neural networks (CNNs), or recurrent neural networks (RNNs) based on the nature of the data. Layers and Nodes: Specify the number of layers, nodes per layer, and activation functions used in the neural network.

4. Training Procedure:

Splitting the Data: Divide the dataset into training, validation, and test sets to evaluate the model's performance effectively.
Loss Function: Define the appropriate loss function for the classification task (e.g., categorical cross-entropy).
Optimization Algorithm:

Select an optimization algorithm (e.g., Adam, SGD) and tune hyperparameters for optimal training.
Regularization Techniques: Apply dropout or other regularization techniques to prevent overfitting.

5. Model Evaluation:

Performance Metrics: Specify the evaluation metrics used (e.g., accuracy, precision, recall, F1-score).
Cross-Validation: Employ cross-validation to ensure robustness and reliability of the model's performance.
Comparison with Baselines: If applicable, compare the neural network model's performance with baseline models or existing methods.

6. Hyperparameter Tuning:

Grid Search or Random Search: Describe the approach used for hyperparameter tuning, optimizing parameters such as learning rate, batch size, and model architecture.

7. Ethical Considerations:

Data Privacy: Address any ethical considerations related to data privacy and confidentiality.

Bias Mitigation: Discuss efforts to identify and mitigate biases in the dataset or model predictions.

8. Software and Hardware:

Programming Language: Specify the programming language (e.g., Python) and the deep learning framework (e.g., TensorFlow, PyTorch) used.
Computational Resources: Mention the hardware specifications used for model training (e.g., GPU, CPU).

9. Validation and Sensitivity Analysis:

Validation Results: Present results from the validation set, demonstrating the model's ability to generalize to new data.
Sensitivity Analysis: Explore the model's sensitivity to changes in hyperparameters or variations in the dataset.

10. Implementation Details:

Code Availability: If applicable, provide information on the availability of the code, facilitating reproducibility.
Documentation: Offer a concise summary of key code components and functions. By following this methodology, we aim to achieve a robust and accurate neural network model capable of effectively classifying molecular subtypes of head and neck cancer based on gene expression levels.

Materials and Methods:

1) Data collection: Source of gene expression data, patient cohort details.

2) Preprocessing steps: Data cleaning, normalization, and feature selection.

3) Description of the neural network architecture chosen.

4) Training procedure: Splitting data, model training, and validation.

Discussion

Building and training a neural network model for classifying molecular subtypes of head and neck cancer involves using gene expression levels as features. The process typically includes data preprocessing, where gene expression data is cleaned and normalized. Then, a neural network architecture is chosen, considering factors like the number of layers and neurons.

During training, the model learns to map gene expression patterns to specific cancer subtypes by adjusting its weights through iterations. Validation datasets help assess the model's performance and prevent overfitting. Hyperparameter tuning may be necessary for optimal results.

The trained model can then be applied to new data for subtype classification, aiding in the understanding and diagnosis of head and neck cancer based on molecular characteristics.

Concluding Remarks

In conclusion, the endeavor to build and train a neural network model for classifying molecular subtypes of head and neck cancer through gene expression levels is a promising and impactful avenue in cancer research. The utilization of advanced machine learning techniques demonstrates potential for more accurate and efficient classification, offering valuable insights into personalized treatment strategies and improving our understanding of the molecular intricacies of this complex disease. However, ongoing validation and refinement of the model are crucial to ensure its reliability in clinical settings and to further advance our capabilities in the field of cancer diagnostics and treatment.

FUTURE WORK:

1.Automated Hyperparameter

- **Explore and implement automated hyperparameter tuning techniques such as grid search, random search, and Bayesian optimization.**

- **Evaluate the feasibility of using tools like Hyperopt, Optuna, or scikit-optimize to automate the hyperparameter tuning process**

2. Advanced Optimization Algorithms:

- Investigate the use of advanced optimization algorithms like genetic algorithms, particle swarm optimization, or simulated annealing for hyperparameter tuning. - Compare the performance of these algorithms with traditional grid search or random search methods.

3. Ensemble Methods:

- **Experiment with ensemble methods for hyperparameter tuning, such as combining results from multiple models with different hyperparameters to achieve better performance.**
- **Explore techniques like stacking or blending to combine the strengths of different models.**

4. Hyperparameter Importance Analysis:

- **Conduct a sensitivity analysis to identify the most influential hyperparameters on model performance.**
- **Prioritize tuning efforts based on the importance of hyperparameters to achieve more efficient and targeted optimization.**

REFERENCES

1. Deep Learning Books:
- "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
- "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron.

2. Neural Network Architectures:

- For gene expression data, consider looking into literature that explores architectures suitable for bioinformatics, such as “Bioinformatics: Sequence and Genome Analysis” by David W. Mount.

3. Gene Expression Analysis:
- “Introduction to Genomics” by Arthur M. Lesk.
- “Bioinformatics Data Skills” by Vince Buffalo.

4. Deep Learning Framework Documentation:
- TensorFlow: https://www.tensorflow.org/
- Keras: https://keras.io/