Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Gene Expression Based Cancer Detection using Biological Attention Hybrid Framework

Dr. Kakoli Banerjee
Computer Science and Engineering JSS Academy Of Technical Education Noida, India

Sakshi Srivastava
Computer Science and Engineering JSS Academy Of Technical Education Noida, India

Anushka Kandwal
Computer Science and Engineering JSS Academy Of Technical Education Noida, India

Khushi Sharma
Computer Science and Engineering JSS Academy Of Technical Education Noida, India

***Abstract*— The early and correct detection of various types of cancer are based on gene expression data which is currently considered as an essential issue with great impact in bioinformatics research and development. The gene expression data is considered to be of high dimensionality and noisy. However, there exist various challenges to already existing ML based classification methods. In this paper, we propose a Biological Attention Hybrid (BAH) model, a dual-path framework that combines attention-based feature combining both classical and deep learning methodology. This model initially uses an attention mechanism to focus on the most important genes and then processes the data in two different paths. One path uses PCA and SVM, while the other uses a CNN-LSTM model to capture both short and long term patterns in the dataset. At the end, the outputs from both paths are combined using a fusion layer to improve overall performance. We tested the model on the TcgaTargetGtex dataset, and it achieved high accuracy with consistent results with the help of cross-validation. The model gives better performance than traditional methods and also highlights important genes. So, this approach can be used for building more accurate and reliable cancer detection systems in future.**

Keywords— gene expression, cancer classification, attention-based feature selection, PCA, SVM, CNN-LSTM, ensemble learning, dimensionality reduction, machine learning, bioinformatics.

I. INTRODUCTION

According to the World Health Organization, cancer is still one of the leading causes of death globally, responsible for almost one out of every six fatalities (WHO, 2024). Enhancing survival rates, targeted therapies, and lowering healthcare costs all depend on early and accurate detection of cancer. The fast growth in genomic technologies in recent years has made it possible for researchers to examine gene expression profiles. These profiles offer important insights into the molecular processes behind the development of cancer by continuously capturing the activity of thousands of genes. The gene expression data are high dimensional and comprise only a small number of samples but tens of thousands of features. Earlier classification algorithms suffer from overfitting, noise and redundancy. Hence dimensionality reduction is crucial for getting the most important features while removing irrelevant genes.

Various dimensionality reduction techniques like PCA have proven to be one of the most effective unsupervised methods. PCA converts high dimensional data into a reduced components in order to maximize the variance of original dataset. Support Vector Machines is the most important supervised learning algorithms for bioinformatics applications in the of classification. SVM manage non linear and high dimensional datasets by making optimal decision boundaries between classes. The combination of PCA with the SVM classifier increases noise reduction and generalization which leads to increased accuracy and computational efficiency.

Earlier methods to cancer classification based on gene expression have mostly followed two methodologies: (i) applying PCA and SVM (ii) deep learning models. While the PCA–SVM method are efficient but deep learning models are capable of finding complex non-linear gene patterns. But no approach alone can fully study the biological structure of genes data.

An important but not yet explored challenge in this domain is finding biologically important genes from many genes in the dataset. All genes do not contribute equally to cancer classification some have more impact than others. Finding important genes can increase the performance and reliability of model. Attention mechanism is used find important genes by learning gene importance weights during training.

PCA–SVM based frameworks showed strong performance, however various essential limitations remained. Gene relevance for the classification task relied on unsupervised dimensionality reduction, which does not specifically take account of relevance in the gene dataset. As a result, biologically important genes may not be given desired importance.

PCA converts the data into linear combinations of genes, that helps in limiting interpretability at the gene level. Already known algorithms do not catch extensive non linear gene expressions and long term dependencies that are present in gene expression data.

The deep learning methodologies points out few of the challenges by understanding non linear expressions but they mostly show overfitting due to the small dataset size and high dimensional of genes data. Also, most of the already existing methods consider machine learning and deep learning as separate models and do not combine them.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

So, there is a need to introduce a new framework that combines both of them: (i) find biologically important genes (ii) reduce dimensionality of data (iii) capture patterns in genes data (iv) combine multiple learning methods into one.

So, for gene expression based cancer detection this paper introduces a novel framework called Biological Attention Hybrid (BAH) model. This model combines attention mechanism with both machine learning and deep learning approaches.

The model consists of three layers:

1. Attention layer which assigns importance weights to genes using attention mechanism.
2. A dual path layer where one path is PCA followed by an SVM and the second is CNN-LSTM to capture both short and long range patterns in gene expression.
3. A fusion layer which combines the output from both layers.

This study tries to improve older methods (like PCA and SVM) by adding new techniques such as attention mechanism and deep learning models. Due of this the model becomes more accurate and easier to understand. It can show which genes are important for cancer detection and make the results with real medical meaning.

II. Related Work/ Literature Review

A. Foundational and survey studies

Prior research showed that high dimensional gene expression data can identify disease categories and develop predictive models, hereby creating the experimental framework of feature selection combined with classification for molecular diagnostics. Detailed reviews on computational learning for gene expression data and support vector machines in medicine gives solution of methodologies and practical considerations, including feature selection, normalization, and cross validation. These survey articles highlight that SVM continues to be a reliable option for small sample, high dimensional issues, particularly when combined with meticulous preprocessing and dimensionality reduction.

B. PCA as a fundamental tool for dimensionality reduction

PCA is extensively utilized in gene expression research for noise reduction, visualization, and as a preprocessing measure prior to classification. Several recent papers and reviews reinforce that PCA is computationally efficient, scales well to genome scale datasets, and often outperforms or complements more complex hidden variable methods for large transcriptomic datasets. Notably, Zhou et al. (2022) showed PCA's practical advantage in large genomics analyses, arguing it is fast and effective for removing major confounders and summarizing variance

C. VM: strengths and modern variants

Because of their margin based generalization and capacity to function in extremely high dimensions, Support Vector Machines (SVMs) have been routinely validated for microarray and RNA seq. classification tasks. Numerous SVM variations and kernel strategies (kernel selection, scalable solvers, ensemble SVMs) have been modified for biomedical applications, according to recent methodological reviews. For instance, current summaries of SVM applications in medicine cover useful advice on handling class imbalance and hyperparameter tuning, which are crucial in tasks involving tumors versus normal tasks.

D. Empirical PCA+SVM papers and case studies

Several applied studies from 2018 to 2025 have implemented PCA in combination with SVM for cancer classification using microarray and RNA-seq datasets. These works evaluate different preprocessing strategies, dimensionality reduction pipelines, and classifier optimizations. A concise summary of notable recent studies is presented below.

- **2024 – S. Al Azani et al.**

  Dataset: RNA-seq. (TCGA)

  Method: Data filtering → PCA → machine-learning models including SVM

  Performance: Addresses the curse of dimensionality and demonstrates that PCA combined with SVM achieves strong accuracy on benchmark datasets.

- **2024 – R. Van et al.**

  Dataset: Multiple RNA-seq. datasets

  Method: Comparison of preprocessing pipelines followed by machine-learning classification

  Performance: Shows that normalization and preprocessing choices significantly influence PCA outputs and SVM performance.

- **2024 – A. Razzaque et al.**

  Dataset: Leukemia, colon, and prostate microarray datasets

  Method: PCA → Particle Swarm Optimization (PSO) for component selection → SVM

  Performance: PCA reduces computational time, and PSO helps optimize principal component selection. The combined PCA–PSO–SVM framework achieves strong classification accuracy.

- **2023 – F. Alharbi et al.**

  Dataset: Survey of multiple gene-expression studies

  Method: Machine learning approaches for gene-expression classification

  Performance: SVM combined with dimensionality reduction or feature selection techniques as a consistently strong baseline for cancer classification.

- **2022 – H. J. Zhou et al.**

  Dataset: Genomics datasets

  Method: PCA versus hidden variable inference methods evaluation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Performance: Shows that PCA is efficient and often outperforms more complex variable.

E. Comparative insights and methodological lessons

Many conclusions can be drawn from the recent literature:

1. Batch correction, low variance gene filtering, and RNA seq. normalization (TPM/RPKM/log transforms) have a significant impact on PCA components and the performance of the resulting SVM. This sensitivity is brought to light by studies comparing RNA seq. preprocessing pipelines, which advise clear disclosure of preprocessing decisions.
2. Using PCA to compress the feature space before SVM reduces computational cost and mitigates overfitting. Several empirical studies report faster training and improved or at least competitive accuracy. PCA frequently eliminates technical variation and noise from very high dimensional data that would otherwise confound classifiers.
3. The Trade-off between interpretability and variance capture. Since PCA components are linear combinations of numerous genes, biological interpretation of them may be challenging because gene level contributions can be ambiguous. Therefore, in order to maintain interpretability, a number of papers combine PCA with supervised feature selection or post hoc analysis of loadings.
4. Hybrid and optimized pipelines outperformed naïve baselines. Generally, works that optimize component selection via heuristic search (PSO, genetic algorithms), combine PCA with wrapper methods (e.g., SVM RFE), or adjust the number of principal components report better classification than PCA→SVM with an arbitrary K.
5. Emerging trend: deep methods and multi omics, but SVM is still a solid foundation. Although stacked/ensemble models and deep learning and multi omics integration are becoming more popular, numerous reviews warn that traditional pipelines like PCA+SVM are still competitive, simpler to understand, and require fewer resources for small sample sizes

F. Limitations and gaps in the literature

- **Heterogeneous reporting**: A direct comparison of reported accuracies is unreliable due to the fact that many studies do not report the same preprocessing procedures, cross validation techniques, or sample counts.
- **Limited biological interpretability**: Without further analysis (loadings, enrichment), it is difficult to link PCA components to particular biological pathways. After PCA, few studies systematically incorporate pathway analysis.
- **Lack of standardized benchmarks**: In contrast to certain machine learning domains, genomics does not have widely accepted benchmark splits; numerous studies employ disparate cancer types and datasets, which makes meta-analysis challenging.
- **Underexplored supervised dimensionality reduction**: although unsupervised PCA is widely used, supervised techniques (PLS, LDA, supervised PCA variants) and kernel/PCA variants should be more carefully compared to PCA+SVM for RNA seq data.

G. akeaways for this review

We draw the conclusion from the reviewed literature that PCA + SVM remains a viable, competitive and interpretable baseline for gene expression-based cancer classification, particularly when sample sizes and computational resources are constrained. Standardized reporting for reproducibility, thorough preprocessing, methodical hyperparameter tuning, and measures to enhance biological interpretability (gene loading analysis, pathway enrichment of components) should be the focus of future research.

III. Methodological Framework

A. Problem Formulation

Let $X \in \mathbb{R}^{N \times G}$ be a gene expression matrix where:

- N = number of samples
- G = number of genes

$x_i \in \mathbb{R}^G$ shows the gene expression of the patient and $y_i \in \{0,1\}$ denotes the class label (0 = Normal, 1 = Cancer).

The aim of this is to learn a classifier

$$f: \mathbb{R}^G \rightarrow \{0,1\}$$

that reduces the error while finding biologically important genes.

B. Workflow Overview

The general workflow for gene expression–based cancer detection using PCA and SVM involves several stages:

1. **Data Acquisition:**

   Gene expression datasets are collected from public repositories such as the **Gene Expression Omnibus (GEO)** or **The Cancer Genome Atlas (TCGA)**.

   These datasets typically contain samples from both **cancerous and non cancerous tissues.**

2. **Data Preprocessing:**

   Before applying PCA, raw gene expression data undergo:

   - Normalization
   - Handling missing values
   - Noise reduction
   - Feature scaling (important for PCA)
   - Low-expression genes are removed
   - Top high-variance genes are selected

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

3. Stage 1: Attention-Based Feature Weighting

Other than traditional PCA–SVM pipelines, the BAH model introduces an attention mechanism to identify important genes.

Each gene is assigned a weight:

$$A_j = \text{softmax}(z_j/\tau) \times n_{genes}$$

The attention weighted gene vector is:

$$\tilde{x}_i = x_i \odot A$$

This step:

- Give importance to biologically important genes
- reduce noise and irrelevant features

4. Stage 2: Dual-Path Feature Learning

After attention weighting, the data goes through two parallel paths.

Path 1: PCA + SVM

Dimensionality reduction using PCA:

$$Z = XW$$

where:

- X= attention-weighted gene matrix
- W= eigenvectors

Classification using SVM:

$$f(x) = \text{sign}\left(\sum_i \alpha_i y_i K(x_i, x) + b\right)$$

This path:

- reduces dimensionality
- provides uniformness

Path 2: CNN-LSTM (New Addition)

The attention weighted gene sequence is a 1D signal:

- CNN layers are used for finding **local gene interactions**
- LSTM is used for finding **long range dependencies**

This allows the model to learn:

- non-linear relationships
- sequential genomic patterns

5. Stage 3: Fusion via Meta-Learning

Outputs from both paths are combined using a stacking approach:

$$F = [p^{SVM}, p^{CNN}]$$

Final prediction:

$$\hat{y} = \sigma(w_1 p^{SVM} + w_2 p^{CNN} + b)$$

This step:

- integrates complementary information
- improves overall accuracy
- reduces model bias

6. Model Evaluation:

Performance is assessed using metrics such as:

- **Accuracy**: Overall correctness of the classifier. Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- **Precision:** Proportion of positive (cancer) predictions that are correct,

  Precision= $\frac{TP}{TP+FP}$
- **Recall (Sensitivity):** Proportion of actual positives (cancer samples) correctly identified,

  Recall= $\frac{TP}{TP+FN}$
- **F1 Score:** Harmonic mean of precision and recall

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

7. Interpretation & Biological Validation:

Unlike traditional PCA based approaches:

- Attention weights directly highlight important genes
- Top weighted genes can be analysed for biological relevance

This improves:

- interpretability
- clinical usefulness

IV.IMPLEMENTATION

A. D a Acquisition

1. The gene expression dataset **TcgaTargetGtex_rsem_gene_tpm.gz** was stored on Google Drive.
2. It was accessed through Google Colab using the mounted drive.
3. To maintain computational feasibility while preserving biological representation:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- 11,000 genes (feature)
- 5,758 labeled samples (after filtering) were used.

B. Data Preparation

1. The original column containing gene IDs was renamed to gene-id.
2. The dataset was transposed, resulting in:
   - Samples → rows
   - Genes → columns
3. The transposed index was reset and labeled as **Sample**.
4. Class labels were assigned using sample naming conventions:
   - Samples starting with **"GTEX" → Normal**
   - All others → Cancer
5. The final dataset included:
   - Sample (ID),
   - gene expression features,
   - Label (Cancer/Normal).

C. Label Encoding

1. Class labels (Cancer, Normal) were converted into numeric form using label encoder
2. This enabled training with machine learning and deep learning models which need numeric input.

D. Feature Engineering and Preprocessing

1. Gene expression values were standardised using standard scalar: Mean = 0 Standard deviation = 1
2. Low expression genes were removed (genes with negligible signal in >90% samples).
3. The top 3,000 genes were selected based on variance from the filtered genes,
4. Data split: 80% training 20% testing

E. Attention-Based Feature Weighting

1. Attention mechanism was used to learn importance weights for each gene.
2. The attention network includes:
   • D nse layer (512 units)
   • D opout (0.2–0.3)
   • Outpu layer (3,000 weights)
3. Softmax normalization with temperature scaling was used in order to predict attention scores.
4. The weighted gene matrix was formulated using element multiplication:

$$\tilde{x} = x \odot A$$

5. Hyperparameter tuning was performed on 27 configurations and the best model was selected based on accuracy of validation.

F. Dimensionality Reduction with PCA (Path 1)

1. PCA was used on the attention-weighted gene matrix.
2. Components were selected in order to preserve 90% of total variance.
3. 8 principal parameters were required. This reduced the dimensionality from 3,000 to 8.
4. PCA resulted in a small and noiseless representation for dataset.

G. odel Training Using SVM (Path 1)

1. An SVM classifier was implemented with:
   • K rnel: Radial Basis Function (RBF)
   • = 1
   • gamma = "scale"
2. The model was trained using the 50-dimensional PCA features from the training set.

H. CN LSTM Model (Path 2)

1. The shape of attention weighted gene vectors was changed to sequence format: (samples, 3000, 1)
2. CNN layers obtained the local gene interaction patterns:
   • onv1D (64 filters) → BatchNorm → MaxPool
   • onv1D (32 filters) → BatchNorm → MaxPool
3. A Bidirectional LSTM obtained long term dependencies in gene expressions.
4. The final dense layer resulted in classification probabilities.
5. Training used:
   • Opti izer: Adam
   • oss: Binary cross entropy
   • arly stopping in order to prevent overfitting

I. Fusion via Meta-Learning

1. Results from SVM and CNN-LSTM were combined with the help of stacking.
2. Out of fold cross validation was used to avoid accidental data leakage.
3. A Logistic Regression model was used as the meta learner.
4. Final result was obtained using the following:

$$\hat{y} = \sigma(w_1 p_{SVM} + w_2 p_{CNN} + b)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

J. sualization

1. PCA Variance Plot (Displays cumulative explained variance)
2. Confusion Matrix (Displays classification performance)
3. ROC Curve (Analyses model discrimination ability)
4. Attention Weight Distribution (Chooses important genes)

K. odel Evaluation

Performance was evaluated using:

- Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision

$$Precision = \frac{TP}{TP + FP}$$

- Recall

$$Recall = \frac{TP}{TP + FN}$$

- F1 Score

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

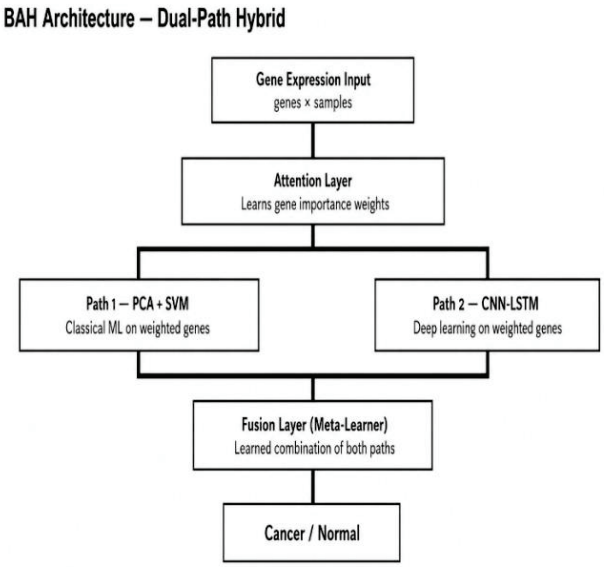
- AUC-ROC

L. Why BAH Works Better

Challenge	Attention Role	PCA + SVM Role	CNN-LSTM Role	Combined Effect
High dimensional data	Selects important genes	Reduces dimensionality	Learns patterns	Efficient learning
Noisy features	Suppresses noise	Removes redundancy	Learns robust features	Better generalization
Small sample size	Focuses on key genes	Works well with low data	Regularized learning	Reduced overfitting
Non linear relations	Highlights gene importance	Linear compression	Captures non linearity	Improved accuracy

C. ypical Workflow Diagram

A flow of the process:



V. Challenges

Biological Attention Hybrid (BAH) frameworks have achieved great results in detection of cancer using this dataset. But there are several technical, practical challenges that limit their full potential.

1. Hi Dimensional and Noisy Data:

- The datasets contain thousands of genes but only hundreds of samples. This will cause overfitting and poor generalization.
- PCA reduces dimensionality but it may ignore important biologically variations if they are not tuned accordingly.

2. M Crash on Full Dataset

30,000 genes and 11,000 samples crashes Google Colab.

3. Hi sensitivity to preprocessing:

Minimal changes in normalization, batch creation or filtering can affect PCA components and SVM performance. This makes results unstable and inaccurate.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

4. mbalanced Datasets:

• n many TCGA datasets, the number of cancerous

Samples are way higher than non cancerous ones.

• This imbalance creates a biasness in the classifier, this results in biased predictions toward the more frequently occurring classes.

VI. Conclusion

The analysis in this paper displays the importance of machine learning in gene expression based analysis for early cancer detection. Conventional methodologies like PCA combined with SVM have already shown good performance for binary as well as multiclass classification. But they still have certain limitations, when it comes to finding complex biological patterns and understanding which genes are actually important.

In this work, we proposed the Biological Attention Hybrid (BAH) model, that enhances the conventional PCA–SVM approach with the help of addition of an attention mechanism and a dual path structure. One path uses PCA and SVM, while the other uses a CNN-LSTM model. This collaboration helps the model in handling high dimensional and noisy gene expression data more effectively and efficiently.

We tested our model on the TcgaTargetGtex dataset, which contains 5,758 samples and 11,000 genes. The model was able to achieve high accuracy with consistent performance with the help of cross validation. This shows that combining attention with hybrid learning models can perform better than conventional PCA–SVM methodologies.

With the help of attention mechanism, we were able to determine 217 important genes, which can act as promising contributors.

However, a few challenges exist, such as class imbalance, computational cost, and validating the important genes. In the future, this work can be extended to multi-class cancer classification and addition of explainable AI techniques for more promising clinical use.

The BAH model displays that combining attention, machine learning, and deep learning together in a single framework can result in both high accuracy and interpretability. This makes it a strong step toward building better and more reliable cancer diagnosis systems in medicine world.

VII. Refrences

[1] T. R. Golub, D. K. Slonim, P. Tamayo, et al., “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, no. 5439, pp. 531 537, 1999.

[2] F. Alharbi and A. Vakanski, “Machine learning methods for cancer classification using gene expression data: A review,” *Bioengineering (Basel)*, vol. 10, no. 2, Art. 173, Jan. 2023. doi: 10.3390/bioengineering10020173

[3] S. Al Azani, O. S. Alkhnbashi, E. Ramadan and M. Alfarraj, “Gene expression-based cancer classification for handling the class imbalance problem and curse of dimensionality,” *International Journal of Molecular Sciences*, vol. 25, no. 4, Art. 2102, Feb. 2024. doi: 10.3390/ijms25042102

[4] . Mani and H. Rajaguru, “A framework for performance enhancement of classifiers in detection of prostate cancer from microarray gene expression tasks,” *Heliyon*, vol. 10, no. 9, e29630, 2024. doi: 10.1016/j.heliyon.2024.e29630

[5] N. Tabassum, S. Islam, S. Rizwan, M. Sobhan, T. A. Chowdhury and S. Ahmed, “Cancer classification from gene expression using ensemble learning and dimensionality reduction,” *Genes*, vol. 15, no. 2, Art. 405, 2024. doi: 10.3390/genes15020405

[6] E. Elhaik, “Principal component analyses (PCA) based findings in genetics & genomics,” *Scientific Reports*, vol. 12, Art. 2369, 2022. doi: 10.1038/s41598 022 14395 4

[7] M. Greenacre, P. J. F. Groenen and T. Hastie, “Principal component analysis,” *Nature Reviews Methods Primers*, vol. 2, Art. 184, 2022. doi: 10.1038/s43586 022 00184 w

[8] E. H. Houssein, Z. Abohashima, M. Elhoseny and W. M. Mohamed, “An efficient binary Harris Hawks Optimization based on quantum SVM for cancer classification tasks,” arXiv:2202.11899 [cs.LG], Feb. 2022.

[9] J. Brown, et al., “Support vector machine classification and validation of cancer tissue microarray gene expression data,” *Bioinformatics*, vol. 16, no. 10, pp. 906 914, 2000.

[10] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.