Deep Learning-Based Cancer Classification from DNA Sequences: Prediction using End-to-End Neural Networks without feature selection

Mythreya
1st Year Grad Student
University of Maryland
mythreya@terpmail.umd.edu

***Abstract*—In this study, we investigate the application of deep learning models for cancer classification based on DNA sequences, eliminating the need for feature selection. Our exploration encompasses five distinct models: a hybrid CNN-RNN, LSTM, biLSTM, k-NN, and k-NN with PCA. The evaluation centers on their efficacy in accurately predicting cancer from DNA samples, emphasizing the potential of end-to-end neural networks in genomics-based medical applications. Our methodology involves a meticulous assessment of each model's performance, focusing on their unique architectural nuances. Importantly, we scrutinize practical implications, considering factors such as model interpretability, computational efficiency, and generalization capabilities. The findings contribute to advancing our understanding of the application of deep learning in genomics-based cancer classification, highlighting avenues for improving medical diagnostics through advanced neural network architectures.**

I. Introduction

Cancer constitutes a significant global health challenge [1], and the timeliness of its detection is pivotal for achieving favorable treatment outcomes [2]. However, existing screening methods face challenges in terms of accuracy, invasiveness, and financial implications. The evolving field of genomics, focusing on genetic analysis, holds promise for transformative advancements in cancer detection and treatment [3]. This project seeks to develop an advanced predictive model utilizing deep learning algorithms to analyse genetic sequence of every individual to predict their susceptibility to cancer.

The approach includes systematically exploring various deep learning models while intentionally avoiding complicated feature engineering. At the same time, a carefully curated dataset is being compiled, encompassing extensive cancer-related genetic data, including genetic sequences, clinical features, and pertinent patient information. Stringent preprocessing procedures are put in place to guarantee the high quality and suitability of the data for future model training.

Following the preprocessing phase, the predictive model undergoes a methodical training regimen using the refined dataset. Subsequent evaluations encompass a comprehensive assessment of its performance metrics and its capacity to generalize beyond the confines of the training set. The resulting predictive efficacy of the model becomes the focal point of our inquiry, prompting a detailed analysis to elucidate the underlying determinants steering its prognostications.

After preprocessing, the predictive model goes through systematic training using the improved dataset. Further evaluations involve a thorough assessment of its performance metrics and its ability to generalize beyond the training set. The resulting predictive effectiveness of the model becomes the main focus of our investigation, leading to a detailed analysis to clarify the factors guiding its predictions.

II. Related Work

Researchers have studied the application of DNA methylation data for predicting different cancer types such as breast, colon, head, kidney, lung, thyroid, and uterine cancer [4]. To tackle the issues of high-dimensional data and noise in DNA methylation, a hybrid approach that combines feature selection and extraction methods was used, yielding favorable outcomes with various classification models. [5].

The second paper aims to comprehensively outline the genetic basis of prostate cancer, utilizing Genome-Wide Association Studies (GWAS) for common genetic variant identification and Next-Generation Sequencing (NGS) to pinpoint rare variants. It generates Polygenic Risk Scores to estimate genetic risk for prostate cancer [6]. While the paper contributes valuable insights into genetic foundations and underscores the importance of understanding biological mechanisms and clinical implications, limitations include the potential incompleteness of GWAS in explaining heritability, challenges in NGS bioinformatics analysis, and complexities in interpreting functional studies, as well as potential representativeness issues in family studies [6].

The next paper focuses on using a machine learning-based approach to predict the tissue-of-origin of cancer through somatic mutation data, employing a random forest classifier developed from 4,000 tumors across 24 cancer types. The classifier, trained with 220 informative somatic mutations, achieved an impressive 88% average accuracy in tissue-of-origin prediction [7]. Notably, the model's advantage lies in

its ability to identify the tissue-of-origin for metastatic tumors with unknown primary sites. However, limitations include the exclusion of non-coding mutations and the necessity for further validation in larger cohorts [7].

Another paper presents a deep learning-based method for predicting relapse in prostate cancer, leveraging multi-omics data from over 400 patients, including gene expression, DNA methylation, and miRNA profiles. Through the integration of multiple omics data, the study aims to provide a comprehensive understanding of biological mechanisms related to relapse. The model's performance, assessed through cross-validation and survival analysis, yielded a notable 76.6% accuracy in predicting relapse. The advantages of this approach include its integration of multiple omics for a holistic understanding of biological mechanisms, high accuracy in relapse prediction, and its potential as a valuable tool for personalized treatment decisions and patient management. However, the study acknowledges the need for further validation in larger cohorts and highlights the necessity for improvements in the interpretability of the deep neural network-based feature learning [4]. In another study, the 'moderated t-statistics' method was employed to identify methylation sites that demonstrated the most notable differential methylation expression. Subsequently, the variance was mitigated using the Empirical Bayes approach, and p-values were adjusted using the Benjamini–Hochberg procedure. Furthermore, to model the complex relationships within the data, two multi-layer feedforward neural networks were constructed. Each network comprised an input layer, multiple hidden layers, and an output layer. This architecture was designed to capture the intricate patterns of methylation expression and contribute to a comprehensive understanding of the underlying biological processes. [8] [9] [10] [11] [12] [13]

III. Methodology

A. Dataset

We will primarily use data from 2 publicly available datasets. These are:

1) SBCB
2) GEO

Our research methodology revolves around the utilization of two publicly available datasets, namely SBCB [14] and GEO [15], as the primary sources of our data. The focal point of our investigation rests on extracting crucial insights from these datasets, with a primary emphasis on the DNA sequence.Our main goal is to find patterns and traits linked to different types of cancer within genetic information. Our predictive model looks at all cancer types instead of separating the data by specific type. Our dataset also includes samples from people without cancer, which helps the model identify differences between healthy individuals and those with cancer. We collected a dataset with 4,396 samples, each containing a wide range of 54,675 genes, covering 15 different cancer types, showing the thoroughness of our genetic data research.

B. Model 1: CNN-RNN Hybrid

Our model's architecture is carefully designed, incorporating two main deep learning elements. At its center is the Input Layer, serving as the entry point for processing complex gene sequences. The CNN layer, a crucial section, unfolds in two distinct parts: first, a Convolution layer, followed seamlessly by a max-pooling layer, each contributing to the model's hierarchical feature extraction capabilities. The subsequent layer, an RNN layer, utilizes LSTM cells (Long Short-Term Memory) to effectively recognize sequential patterns, enhancing the model's capacity for nuanced comprehension. The fusion of both the CNN and RNN components is smoothly achieved through the concatenate function, a strategic tool provided by TensorFlow. The model's output layer, a critical point in its decision-making process, consists of a single neuron utilizing the sigmoid activation function, strategically designed for its binary classification nature. The intricacies of the model are encompassed in the compilation phase, where the Adam optimizer plays a central role in guiding the model towards optimal performance. Upon rigorous testing, our model reveals its prowess with an average accuracy of 92.10% on the testing dataset, a testament to its robust learning and predictive capabilities in the intricate domain of genetic sequence analysis. Fig.1 and Fig.2 is a plot of the history of the accuracy and loss of the model.

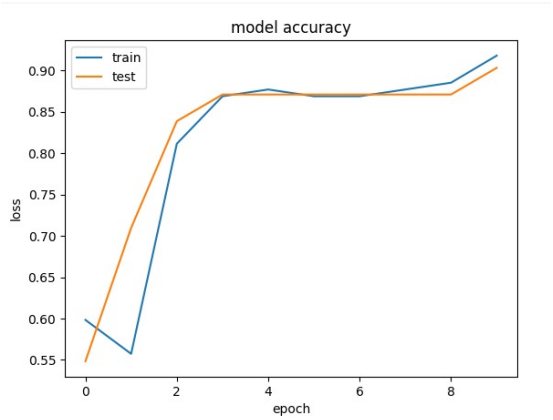

Fig. 1: Hybrid CNN-RNN Model Accuracy

C. Model 2: LSTM

The structural composition of our model unfolds with precision through a four-layer configuration: an input layer, followed by two hidden layers, each intricately woven with 64 LSTM cells, and culminating in an output layer housing a singular dense neuron equipped with the sigmoid activation function. In emulation of its predecessor, the input layer serves as the inaugural gateway for gene sequences, orchestrating their seamless transition into the hidden layers celebrated for their nuanced grasp of sequential pattern recognition. This sophisticated architecture is meticulously synthesized through the application of the Adam optimizer, a pivotal element in refining the model's parameters for optimal performance.

Rigorously tested, our model reveals its analytical prowess, giving an average testing data accuracy of 82.26%, thereby affirming its substantive competence within the intricate realm of genetic sequence analysis. The accuracy and loss graph can be seen in Fig.3 and Fig.4, and they show that the model performance has improved with each epoch.

D. Model 3: biLSTM

In the development of the biLSTM model, the compilation process mirrors that of the LSTM model, diverging solely in the incorporation of bi-directional LSTM cells within the two hidden layers. Despite this nuanced alteration, the discernible impact on the model's performance is relatively moderate, as reflected by an average testing data accuracy of 83.88%. This model however did not perform much better than LSTM, with only a slight improvement in accuracy. The accuracy and loss graph for this model (as seen in Fig.5 and Fig.6) look very similar to the LSTM model (Fig.3 and Fig.4)

E. Model 4: k-NN

The k-NN (k-Nearest Neighbors) model employed in this context was characterized by its simplicity, with a parameter setting of k, representing the number of nearest neighbors, set to 5. This model was directly fitted with the preprocessed data, emphasizing a straightforward approach to pattern recognition and classification. The essence of the k-NN algorithm lies in its reliance on the proximity of data points in feature space, where predictions are determined by the consensus of the k closest neighbors [16]. This model resulted in a testing accuracy of 84.52%.

F. Model 5: k-NN with PCA

In the instantiation of Model 5, a k-NN with PCA, Principal Component Analysis (PCA) played a pivotal role in dimensionality reduction for gene sequences [17]. The transformed gene sequences subsequently served as the training input for the k-NN model. Notably, the k-NN model coupled with PCA demonstrated a remarkable average testing data accuracy of 83.91%. However, surprisingly, k-NN performed much better without PCA.

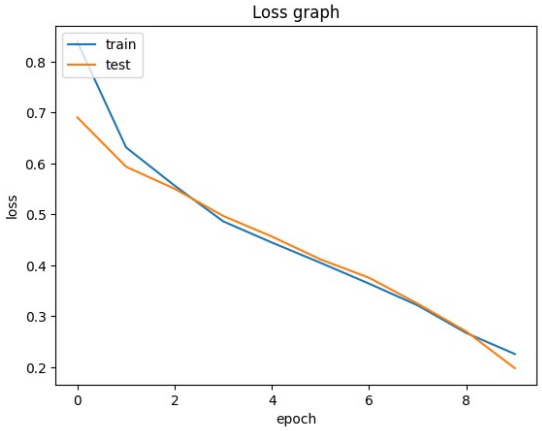

Fig. 2: Hybrid CNN-RNN Loss Graph

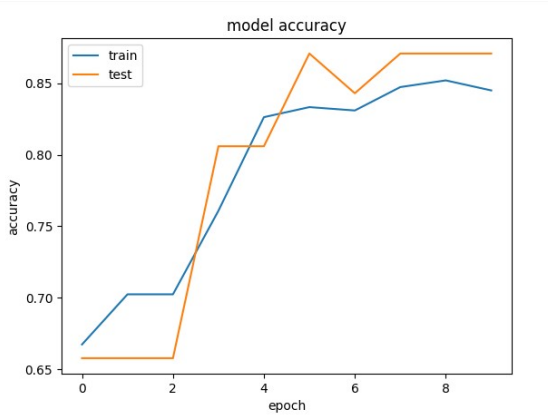

Fig. 3: LSTM Model Accuracy

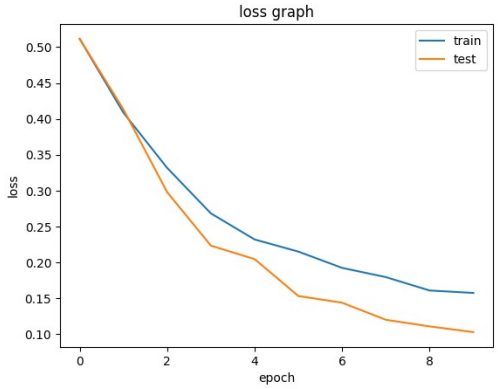

Fig. 4: LSTM Loss Graph

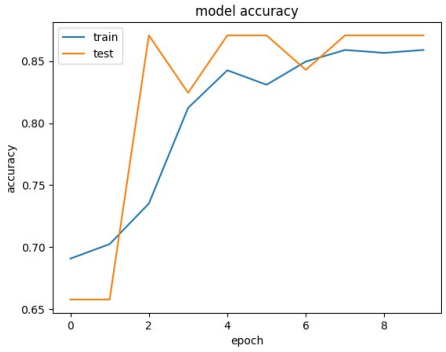

Fig. 5: Bi-LSTM Model Accuracy

IV. Results

The performance evaluation of our cancer prediction models yielded insightful results. The Hybrid CNN RNN Model emerges as the clear frontrunner, showcasing a remarkable accuracy of 92.10% (Fig. 7). Further insights from performance metrics (Fig. 8) reveal a high recall of 0.9294, precision of 0.8895, and an F1 score of 0.9090. This robust performance

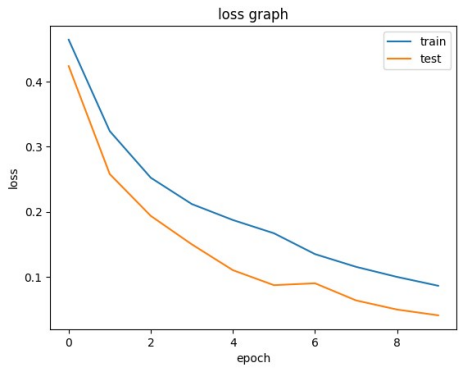

Fig. 6: Bi-LSTM Loss Graph

underscores the model's effectiveness in capturing intricate relationships within genetic sequences, leveraging the synergies of convolutional and recurrent neural network architectures.

TARGET / OUTPUT	Positive	Negative	SUM
Positive	1224 39.47%	152 4.90%	1376 88.95% 11.05%
Negative	93 3.00%	1632 52.63%	1725 94.61% 5.39%
SUM	1317 92.94% 7.06%	1784 91.48% 8.52%	2856 / 3101 92.10% 7.90%

Fig. 7: CNN-RNN Confusion Matrix

Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
Positive	0.8895	0.1105	0.9294	0.0706	0.9090
Negative	0.9461	0.0539	0.9148	0.0852	0.9302
Accuracy	0.9210				
Misclassification Rate	0.0790				
Macro-F1	0.9196				
Weighted-F1	0.9212				

Fig. 8: CNN-RNN Other Performance Metrics

In analyzing the results further, the LSTM Model achieved an accuracy of 82.26%, as illustrated by the confusion matrix (Fig. 9). While this accuracy is respectable, a closer examination reveals lower recall (0.6463) and F1 score (0.7461), suggesting potential challenges in capturing subtle patterns within the genetic data. The LSTM Model's performance might be attributed to its reliance on long short-term memory units, which excel at capturing sequential dependencies. However, the intricate relationships present in genetic sequences may not be fully harnessed by this architecture.

TARGET / OUTPUT	Positive	Negative	SUM
Positive	808.0 26.06%	102.0 3.29%	910 88.79% 11.21%
Negative	448.0 14.45%	1743.0 56.21%	2191 79.55% 20.45%
SUM	1256 64.33% 35.67%	1845 94.47% 5.53%	2551 / 3101 82.26% 17.74%

Fig. 9: LSTM Confusion Matrix

Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
Positive	0.8879	0.1121	0.6433	0.3567	0.7461
Negative	0.7955	0.2045	0.9447	0.0553	0.8637
Accuracy	0.8226				
Misclassification Rate	0.1774				
Macro-F1	0.8049				
Weighted-F1	0.8161				

Fig. 10: LSTM Other Performance Metrics

On the other hand, the Bi-LSTM Model presents a notable improvement with an accuracy of 83.88%. The confusion matrix (Fig. 11) indicates effective classification performance. Notably, the bidirectional aspect of the long short-term memory networks contributes to a higher recall (0.6720), precision (0.9056), and F1 score (0.7715). This bidirectional processing allows the model to capture information from both past and future time steps, enhancing its ability to discern complex patterns in the genetic data. The Bi-LSTM Model's success in achieving a balanced trade-off between precision and recall underscores its suitability for our cancer prediction task.

The KNN Model and the KNN Model with PCA present accuracies of 84.52% and 83.91%, respectively, based on their confusion matrices (Fig. 13 and Fig. 15). These non-neural network approaches exhibit competitive results, emphasizing the importance of exploring diverse model architectures. However, the Hybrid CNN RNN Model's superior accuracy and

TARGET / OUTPUT	Positive	Negative	SUM
Positive	844.0 27.22%	88.0 2.84%	932 90.56% 9.44%
Negative	412.0 13.29%	1757.0 56.66%	2169 81.01% 18.99%
SUM	1256 67.20% 32.80%	1845 95.23% 4.77%	2601 / 3101 83.88% 16.12%

Fig. 11: Bi-LSTM Confusion Matrix

Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
Positive	0.9056	0.0944	0.6720	0.3280	0.7715
Negative	0.8101	0.1899	0.9523	0.0477	0.8754
Accuracy	0.8388				
Misclassification Rate	0.1612				
Macro-F1	0.8235				
Weighted-F1	0.8333				

Fig. 12: Bi-LSTM Other Performance Metrics

performance metrics suggest that the nuanced features captured by neural networks significantly contribute to effective cancer prediction.

TARGET / OUTPUT	Positive	Negative	SUM
Positive	1008 32.51%	180 5.80%	1188 84.85% 15.15%
Negative	300 9.67%	1613 52.02%	1913 84.32% 15.68%
SUM	1308 77.06% 22.94%	1793 89.96% 10.04%	2621 / 3101 84.52% 15.48%

Fig. 13: KNN Confusion Matrix

Upon comparison, the Hybrid CNN RNN Model outperformed all other models, exhibiting the highest accuracy and superior recall, precision, and F1 score. With an accuracy of 92.10%, the Hybrid CNN RNN Model stands as the most effective model for predicting cancer likelihood based on genetic sequences in our study.

Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
Positive	0.8485	0.1515	0.7706	0.2294	0.8077
Negative	0.8432	0.1568	0.8996	0.1004	0.8705
Accuracy	0.8452				
Misclassification Rate	0.1548				
Macro-F1	0.8391				
Weighted-F1	0.8440				

Fig. 14: KNN Other Performance Metrics

TARGET / OUTPUT	Positive	Negative	SUM
Positive	997 32.15%	189 6.09%	1186 84.06% 15.94%
Negative	310 10.00%	1605 51.76%	1915 83.81% 16.19%
SUM	1307 76.28% 23.72%	1794 89.46% 10.54%	2602 / 3101 83.91% 16.09%

Fig. 15: KNN(with PCA) Confusion Matrix

Class Name	Precision	1-Precision	Recall	1-Recall	f1-score
Positive	0.8406	0.1594	0.7628	0.2372	0.7998
Negative	0.8381	0.1619	0.8946	0.1054	0.8655
Accuracy	0.8391				
Misclassification Rate	0.1609				
Macro-F1	0.8327				
Weighted-F1	0.8378				

Fig. 16: KNN(with PCA) Other Performance Metrics

V. Conclusion

In this study, we explored the application of deep learning models for cancer classification from DNA sequences, specif-

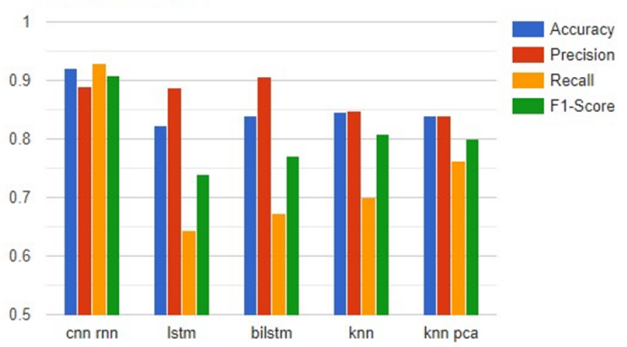

Fig. 17: Overall Analysis

ically investigating a hybrid CNN-RNN, LSTM, biLSTM, k-NN, and k-NN with PCA. The results revealed interesting insights into the performance of these models. Our results show that the k-NN model, a conventional machine learning approach, showed signs of overfitting on the provided dataset. Despite its simplicity, k-NN had difficulty in generalizing well, highlighting the importance of careful parameter tuning and dataset characteristics for traditional models. In contrast, the deep learning models such as hybrid CNN-RNN, LSTM, and biLSTM exhibited strong performance without overfitting. This indicates the capability of deep learning structures to automatically learn complex patterns and representations directly from raw DNA sequences, reducing the risk of overfitting seen in traditional machine learning models like k-NN.

Moreover, our findings provide insight into the importance of feature selection in genomics-based classification tasks. In contrast to conventional machine learning models, the deep learning models in our research did not necessitate explicit feature selection. This underscores the inherent ability of deep learning architectures to automatically identify pertinent features, making traditional feature selection less crucial in the realm of DNA sequence classification.

In summary, our research emphasizes the effectiveness of deep learning models, including the hybrid CNN-RNN, LSTM, and biLSTM, for categorizing cancer from DNA sequences. Unlike traditional models like k-NN, which may encounter overfitting issues, the comprehensive nature of deep learning structures enables them to utilize the inherent complexity of genomic data, potentially minimizing the requirement for manual feature selection. These results add to the increasing understanding of applying deep learning to genomics and offer potential advancements in medical diagnostics.

REFERENCES

[1] WHO, "The top 10 causes of death," 2020.
[2] C. R. UK, "Why is early cancer diagnosis important?" 2023.
[3] M. M. Pomerantz and M. L. Freedman, "The genetics of cancer risk," *Cancer J*, vol. 17, no. 6, pp. 416–422, 2011.
[4] Z. Wei, D. Han, C. Zhang, S. Wang, J. Liu, F. Chao, ..., and G. Chen, "Deep learning-based multi-omics integration robustly predicts relapse in prostate cancer," *Frontiers in Oncology*, vol. 12, p. 1, 2022.
[5] A. A. Raweh, M. Nassef, and A. Badr, "A hybridized feature selection and extraction approach for enhancing cancer prediction based on dna methylation," *IEEE Access*, vol. 6, pp. 15 212–15 223, 2018.
[6] H. Ni Raghallaigh and R. Eeles, "Genetic predisposition to prostate cancer: an update," *Familial Cancer*, vol. 21, pp. 101–114, 2022.
[7] X. Liu, L. Li, L. Peng, B. Wang, J. Lang, Q. Lu, ..., and L. Zhou, "Predicting cancer tissue-of-origin by a machine learning method using dna somatic mutation data," *Frontiers in genetics*, vol. 11, p. 674, 2020.
[8] B. Liu, Y. Liu, X. Pan, M. Li, S. Yang, and S. C. Li, "Dna methylation markers for pan-cancer prediction by deep learning," *Genes*, vol. 10, no. 10, p. 778, 2019.
[9] H. C. Wong, C. S. K. Lee, and D. L. Tong, "Pathway analysis of marker genes for leukemia cancer using enhanced genetic algorithm-neural network (engann)," pp. 118–121, 2018.
[10] L. V. Povoa, U. C. B. Calvi, A. C. Lorena, C. H. C. Ribeiro, and I. T. Da Silva, "A multi-learning training approach for distinguishing low and high risk cancer patients," *IEEE Access*, vol. 9, pp. 115 453–115 465, 2021.
[11] C. Xia, Y. Xiao, J. Wu, X. Zhao, and H. Li, "A convolutional neural network based ensemble method for cancer prediction using dna methylation data," pp. 191–196, 2019.
[12] L. Muflikhah, N. Widodo, and W. F. Mahmudy, "Prediction of liver cancer based on dna sequence using ensemble method," pp. 37–41, 2020.
[13] Y. A. Abass, S. A. Adeshina, N. N. Agwu, and M. M. Boukar, "Analysis of prostate cancer dna sequences using bi-direction long short term memory model," pp. 1–6, 2021.
[14] Sociedade Brasileira de Computação Bioinformática, "Cumida," https://sbcb.inf.ufrgs.br/cumida.
[15] National Center for Biotechnology Information (NCBI), "NCBI GEO," https://www.ncbi.nlm.nih.gov/geo/.
[16] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 1255–1260.
[17] B. M. Salih Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, p. 20–30, Apr. 2021. [Online]. Available: https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032