Development And Application of Artificial Neural Network and Deep Learning Frameworks for Information Processing

Maram Y. Al-Safarini; Rania A. Baashirah

Recommended citation:

Al-Safarini, Maram Y; Baashirah, Rania A (2024). "Development And Application of Artificial Neural Network and Deep Learning Frameworks for Information Processing". *Profesional de la información*, v. 33, n. 6, e330612.

https://doi.org/10.3145/epi.2024.ene.0612

Manuscript received on 11th March 2024
Accepted on 06th October 2024

Maram Y. Al-Safarini ✉
https://orcid.org/0000-0002-4899-4062
Computer Science Department
Zarqa University, Jordan
maram@zu.edu.jo

Rania A. Baashirah
https://orcid.org/0000-0001-6877-6698
Software Engineering Department
University of Business & Technology
Saudi Arabia
r.baashirah@ubt.edu.sa

Abstract

Massive datasets of molecular sequences, medical images, and other structured information pose challenges of their utilization and interpretation through traditional data analysis methods. The current study aims to explore how Artificial Intelligence procedures like deep learning can improve predictive modeling and pattern recognition in healthcare analytics and information processing. This study proposes a two-stage deep learning model that combines long short-term memory (LSTM), convolutional neural networks (CNNs), and natural language processing (NLP) techniques. This combination helps improve the accuracy of predictions The study also proposed the use of "SENIES," the DNA Shape Enhanced Two-Layer Deep Learning Predictor, a computational method used to identify enhancer regions within DNA sequences. The study used a scientific and exploratory methodology for the identification and characterization of the enhancer, utilizing a sample of active enhancers from a cohort of 9,000 cancer patients from a machine learning-powered database. The data was analyzed through Mathew's correlation coefficient principle using steps like precision recall, specificity, and accuracy. This method is commonly used to evaluate categorization accuracy. The study found that when identified enhancers are placed next through AI-based evaluation to assess their characteristics, they can decipher their regulatory functions, and determine their relationships to the target genes. It was also concluded that the proposed models work well with different dataset sizes, making it flexible for various applications, leading to the integration of AI-driven biological data along with the recognition and prioritization of functional non-coding mutations, as an efficient method of cancer research. The study recommends incorporation of some of the additional AI-driven methods for better results and accurate predictions. Future research should focus on integrating predictive models for real-time data analysis, which would help improve the development and effectiveness of such models.

Keywords

Deep Learning, Artificial Intelligence, Neural Network, Information Processing.

1. Introduction

Artificial intelligence (AI) is aiming to transform the healthcare by improving medical decision-making processes, disease detection and treatment (**Rajpurkar *et al.***, 2022). The given advancement in the field of AI has provided benefits to several stakeholders (**Deng *et al.***, 2022; **Rajpurkar *et al.***, 2022). However, they also create new challenges, and as a results, pushing towards the adoption of better systems and structures while using the AI on effective grounds (**Sun; Medaglia**, 2019). Unlike earlier available technologies, AI in healthcare requires a lot of new and innovative ways of managing and overseeing its usage. This is because AI is constantly learning and becoming more independent in the modern world (**Berente *et al.***, 2021). Simultaneously, AI-based systems are often found full of complexities, making them difficult to understand for many people. For example, AI models can change over time due to shifts in their data or real-world conditions, it is quite hard for

the decision-makers such as hospital administrators to observe these technological changes. AI is always evolving, therefore, healthcare industry needs new strategies to monitor and control its use (**Jöhnk *et al.***, 2021). To address such concerns, decision-makers must have access to the right information at the right time as linked with the AI domain. One of the key points is that information is valuable because it helps people make smart decisions. However, sharing AI-related information is often difficult which further determines the information processing layout. This is partly because different experts like doctors, data scientists, and administrators have their diversified level of knowledge related to AI. Since AI in healthcare involves multiple experts working together, strong coordination is essential for success (**Higgins; Madai**, 2020). Therefore, the way AI is developed and used in healthcare depends on how well information is processed.

There is ongoing research to explore how artificial intelligence and deep learning domains can improve predictive modeling and pattern recognition in healthcare analytics and information processing. Big data in healthcare is growing with a rapid speed due to AI-driven approaches. It is leading to the creation of massive datasets that include molecular sequences, medical images, and other structured information. However, effective utilization of this data is very challenging by using the artificial intelligence and relevant facilities. While traditional data analysis methods demand some sort of the rigorous procedures, the idea of deep learning has proven to be highly effective in managing large-scale medical and biological data. The reason is that traditional experimental methods for identifying these elements are somehow slow and not effective in terms of cost factor. In contrast, AI-based computational methods are faster and more efficient. To tackle these challenges, this study presents a two-stage deep learning model that combines long short-term memory (LSTM), convolutional neural networks (CNNs), and natural language processing (NLP) techniques. This combination helps improve the accuracy of predictions. Moreover, the model is designed to work well with different dataset sizes, making it flexible for various applications. The study also proposed the use of "SENIES" (DNA Shape Enhanced Two-Layer Deep Learning Predictor), a computational method used to identify enhancer regions within DNA sequences by leveraging information about the 3D structure and shape of the DNA alongside traditional sequence data, allowing for better prediction of both the presence of enhancers and their strength.

It has also widely been admitted that DL techniques are getting interest from different researchers for processing information, especially in the domains like bioinformatics and computational biology. One of the key reasons for their ongoing success is their ability to manage and interpret large amounts of biological data. Meanwhile, the neural networks help in transforming some raw data into structured information. Therefore, it is making it easier to analyze the available data and information processing. This research focuses on enhancing bioinformatics through deep learning (DL) to identify and predict important regulatory elements in complex datasets for the purpose of information processing. This study offers a quick and accurate DL architecture as a response to the problems listed above. Furthermore, it is imperative to stay abreast of the latest advancements in research in order to maintain a leading position in this rapidly evolving field.

This study is based on the premise that a more efficient method of cancer research would involve the integration of AI-driven biological data along with the recognition and prioritization of functional non-coding mutations. Within the field of AI-driven biomedical research, the examination of cell-specific enhancer activity is a prominent area of study. The accomplishments of deep learning (DL) in AI-based biological research have been examined in the parts that follow. In the current work, AI-driven computational models were employed to investigate enhancer behavior, uncovering multiple unique connections between enhancers and important epigenetic regulatory processes. The objective of the ensuing investigation was to comprehensively classify cancer-causing mutations in patients with acute leukemia. A significant number of recurrent mutations in diverse epigenetic regulators were effectively identified through deep learning-based analysis. Based on the aforementioned studies, it is now possible to assess how AI-powered models can analyze somatically acquired mutations within epigenetic regulators, evaluating their impact on gene expression and enhancer function during tumor evolution. The findings clarify the crucial role that machine learning-driven transcriptional analysis plays in understanding the etiology of cancer-related genetic disorders.

2. Literature Review

Due to recent developments in DNA fragmentation technologies, the AI-driven bioinformatics discipline has produced large amounts of data. Sequences can be translated through a complex process that uses the sequence to generate protein, even though they cannot convey ready-to-use information. The comparison of the produced sequence with known cancer-related datasets facilitates the assessment of protein expression and diagnosis of malignancy (**Ding *et al.***, 2010). The gathering of biological data has made it more challenging to construct a cogent explanation of the genetic causes of cancer. The difficulties in the treatment and prevention of diseases are also closely related to the wide variation in gene expression levels seen in individuals, which includes a variety of characteristics but is only represented by a small number of samples. The likelihood of a successful recovery is inversely connected with how quickly the disease is recognized (**Chen *et al.***, 2021). Features are retrieved in a hierarchical way within the context of a deep learning (DL) architecture, encompassing multiple degrees of nonlinearity.

Single-solution algorithms and population-based algorithms are two main categories under which metaheuristic AI algorithms fall. In the course of the optimization process, the single-solution method only considers one potential solution. As the process goes on, this solution changes and evolves. On the other hand, the population-based approach deals with

the process of optimization by using a randomly selected search agent. Each agent follows a unique strategy to address the optimization problem. By sharing information about the search area and actively collaborating, the agents reduce the risk of getting stuck in local optima. Moreover, it is further expressed that AI-driven metaheuristic algorithms are widely used in decision-making problems. Besides, the idea of exploration involves searching for new potential solutions. The application of a unified optimization technique can successfully handle the issue of Feature Selection in AI-based models (**Brynjolfsson; McAfee**, 2014) and the difficulties posed by binary optimization. Researchers have developed a number of hybrid methods that combine simulated annealing and the Whale Optimization Algorithm (WOA), such as the Genetic Algorithm (GA), Gray Wolf, and Particle Swarm Optimizer (PSO). Additionally, **Moslehi and Haeri** (2020) have suggested a hybrid strategy that combines the filter and wrapper approaches of feature selection (**Brynjolfsson; McAfee**, 2014). It cannot be guaranteed that the FS issue will reveal improved properties. Furthermore, it is impossible to create an optimizer that is capable of solving all optimization issues, according to the No Free Lunch (NFL) theorem in AI optimization (**Ho; Pepyne**, 2002).

Among the different types of deep learning models, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are showing their outstanding potentials. These models are useful for tasks in the form of data extraction, pattern recognition, and in detecting abnormalities. Additionally, the RNNs are particularly good at processing structured data, while CNNs focus on identifying patterns in structured datasets. Such qualities are making them effective for feature extraction and classification. Moreover, the use of long short-term memory (LSTM) and gated recurrent unit (GRU) based models improve the information retention. Similarly, CNNs have also some sort of potential in dealing with the genetic information. Additionally, DL models help process and organize information by learning from labeled data. Therefore, they are capable enough to classify and predict information, along with generating new data, or remove unnecessary noises in the data and information. However, how well a model learns depends on the specific goal. A key concept in AI is knowledge transfer, which allows a model trained on one task to be adapted for another similar task with minimal extra training. This not only saves time and computing power but also makes artificial intelligence-based models more adaptable. Given the large size of algorithmic datasets, efficient model designs and data management strategies are essential (**Adadi**, 2021). Techniques such as model compression, data augmentation, and mini-batching optimize computational efficiency and resource utilization. In real-time or high-throughput applications, model inference speed is critical, and hardware acceleration using Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs), along with lightweight architectures, can significantly enhance processing efficiency. To assess model performance in AI-driven information analysis, relevant metrics such as precision, recall, F1-score, or the area under the receiver operating characteristic curve (AUC) should be used.

Cancer prediction using AI-driven biological data is a major domain in the current times. AI-powered association studies have found genetic variants linked to an elevated risk of cancer found outside of the coding DNA sequence. These variations are frequently discovered in parts of the DNA where enhancer elements are known to be abundant. Furthermore, the advancement of high-throughput methods has made it possible to precisely identify potential enhancer elements in both tumor and normal cells. In order to make the identification and manipulation of enhancers easier, a variety of AI-driven approaches and analytical techniques have been used in the classification of enhancer activity markers. The specific scenarios associated with it limit the study of enhancers in experimental detection. As enhancers may display activity in particular physiological circumstances while staying inactive in different cell types or states, this procedure necessitates doing numerous trials to find them. Given the significant discrepancies in enhancer predictions produced by current AI-based computational methods, it would be advantageous for the scientific community to provide a thorough analysis of the methodologies and solutions developed in this field. A specific DNA region is examined utilizing a variety of data sources to establish its potential as an enhancer as part of the computational process for discovering enhancers.

Given the aforementioned justifications, a modern, exact, and quick strategy to treat cancer diseases is still required. In cases where the problem is not serious, a rapid and correct diagnosis within a reasonable amount of time is essential. On the other hand, in emergency situations, a prompt diagnosis is essential to reducing the risk to a person's life. Understanding the AI-driven biological predictions generated by DL algorithms is of paramount importance. Interpretability approaches such as saliency maps and attention processes can be employed to unveil the specific segments of a biological sequence that exert influence on the predictions. The selection of model architecture and methodologies will depend on the particular AI-based biomedical task being undertaken and the available data. The field of predicting and analyzing biological sequences with DL is currently undergoing active development in order to enhance efficiency.

3. Research Methodology

The study used a scientific and exploratory methodology for the identification and characterization of the enhancer. Various AI-driven computational methods interpret DNA segments using feature vectors. A list of anticipated enhancers is the main output of the enhancer identification process. In this study, the identified enhancers were placed next through AI-based evaluation to assess their characteristics, decipher their regulatory functions, and determine their relationships to the target genes. Figure 1 shows that to identify enhancers the initial step for data resources is to provide feature vectors that define the dataset. While eRNAs act as feedback to help with the classification of enhancers and the evaluation of their properties, enhancer data is used for screening purposes.

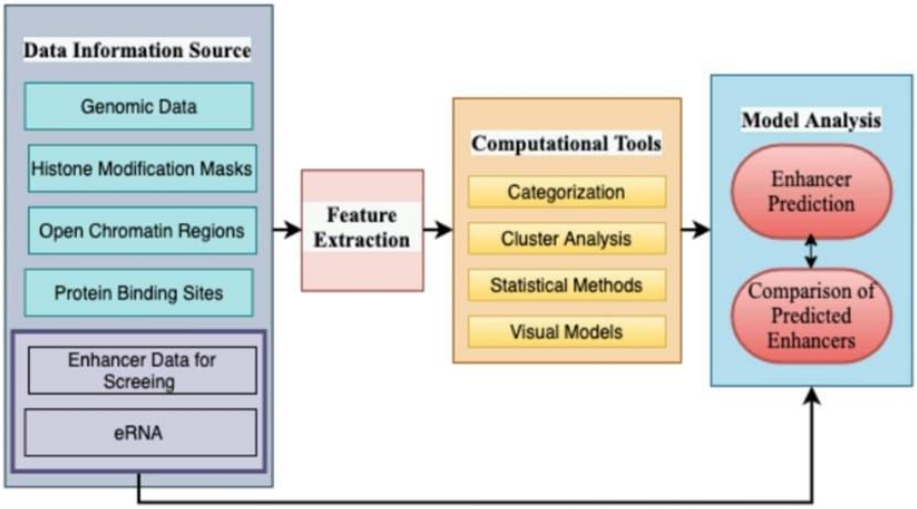


Figure 1: Enhancer Prediction System.

Open chromatin areas, histone modification marks, and protein binding sites are just a few of the markers that have been proposed with varied degrees of success as enhancer predictors. There is still no agreement on a single enhancer identification method that will eventually be used to distinguish and find enhancers within large biological data regions. To enhance and facilitate theoretical techniques for the discovery of enhancers, a wide range of AI-driven mathematical models were developed. Numerous attempts to determine the role of transcriptional factors in cancer have yielded a profound understanding of the mechanisms underlying tumor sustenance and malignant transformation (**Bakrania *et al.***, 2023). An investigation using the distribution of histone marks and RNA-Seq data from the TCGA database yielded a total of 15,808 enhancers. The expression of enhancer RNA was thought to be a sign of enhancer activity. Multiple connections between enhancers and oncogenes have been made, and a comparison between somatic copy number changes and global enhancer interactions has been conducted. Based on the knowledge that the chromatin state significantly affects genetic regulatory patterns, this comparison was made. In terms of sample, this study is the first to thoroughly examine the number of active enhancers in a cohort of 9,000 cancer patients with a variety of cancer types. It also seeks to create a machine learning-powered database that can be used for future research into the underlying mechanisms driving chromatin modifications that aid in the development of cancer. Table 1 provides an in-depth analysis of databases used for each tool enhancer with their advantages and limitations.

Table 1: Sample of the Study.

Tool for Enhancer Prediction	Database Used	Advantages	Limitations
kmer-SVM (**Wang *et al.***, 2023)	EP300/CREPP scores binding sites true positive	50% of predicted enhancers with SVM score above 1.0 are	Sample bias with training data
gkm-SVM (**Velichko *et al.***, 2023)	GN12878 from gene expression omnibus	Improvement in classification accuracy for long TFBS	Overfitting curves
GMFR-CNN (**Das; Toraman**, 2022)	10 ChIp - SEQ Datasets from deep bind	98% prediction accuracy with high precision (98%) & sensitivity (98%) & optimal F measure	Selection bias during training
DeepBind (**Li *et al.***, 2021)	PBM, SELEX ChIp & CLIP seq	Fully automated good score for in vivo data after training from in vitro data	Poor calibration setting in training
Basset (**Painuli *et al.***, 2022)	164 cell types by DNase-seq	Provides a better view of region with high resolution & high prediction accuracy	High computing cost
DeepSea (**Cheng *et al.***, 2022)	Chromatin profiles from ENCODE & Road map data	High prioritisation of eQTL variants, Chromatin effects prediction 0.896, & Transfer learning framework	Unbalanced PR-AUC metric & Lack of reoccurring modelling
DanQ (**Sapoval *et al.***, 2022)	Chromatin profiles from ENCODE & road map data	Performed better than DeepSEA for 97.6% targets	Perfect fit of architecture missing & Model is not fully recurrent, so it cannot process sequences of all lengths
TFImpute (**Yin *et al.***, 2022)	ChIp-seq from ENCODE	Better area under the curve & recall rate in comparison with DeepSEA & gkm-SVM	No linear dependencies of enhancer on transcriptional binding factors. & Best fitting not achieved with all enhancer marks
CSI-ANN (**Durge *et al.***, 2022)	Hela ENCODE (in terms of delay values)	Performed better than profile-based approach & HMM-based chromium, high sensitivity of 84%	Missing statistical depth as only energy & mean are calculated also no non-linear feature used
RFECS (**Lee *et al.***, 2022)	Human embryonic stem cell 24 chromatin modification of primary lung fibroblast cells	Best validation rate of 70% & Misclassification of less than 7%	Worked accurately only for specified cell types in comparison with CSI-ANN
REPTILE (**Lee; Jang**, 2022)	6 Histone modification marks, epigenetic marks & VISTA enhancer browser	High power of prediction (92%) in case of different cell types & better performance than RFECS	Few negatively stated enhancers exhibit enhancer like properties
EMERGE (**Hanczar *et al.***, 2022)	ENCODE, ChIp-seq & NARROWPEAK	Research friendly tool shows the overlap of various datasets	Overlapping sometimes gives misinterpreted dataset true negative values
EnhancerFinder (**Tran *et al.***, 2023)	VISTA enhancer database	Outperformed evolutionary conservation DNA motive in terms of ROC with high value	Prediction are based on single dataset so, the response of method to other known enhancer is unpredictable
Semi-supervised training method (**Hou *et al.***, 2023)	VISTA enhancer database, HICAP enhancer, & FATOM5 enhancer	Outperformed simplified semi-controlled learning with ROC of 0.84	Classifier preparation involves a collection of known positive & negative observations
Pan-cancer analysis (**Gao *et al.***, 2023)	Tumour samples from 33 types of cancer with TCGA RNA-seq dataset	Complete and accurate system for detection of long non-coding RNA & cancerous genes, & This method purposed a novel method for enhancer identification	Different tissues combined due to limited size of data results in increased false positives

The reliability or authenticity of the enhancers sampled for this study is not assured by the use of computational methodologies in enhancer discovery. It is currently difficult to guarantee and confirm the accuracy of identified enhancers because there is not a sufficiently broad and empirically tested collection of enhancer data for AI-driven biological analysis. In the field of enhancer prediction, a significant barrier is the lack of consistent consensus across many approaches. There is a need for thorough AI-powered methodologies that can unquestionably evaluate enhancer activity appropriately. The examination of enhancer element mutations' effects on tumor growth can be facilitated by thorough analysis and clinical testing of these mutations. The use of enhancer sequences that have undergone experimental validation has the potential to increase the effectiveness of cancer therapy. The dataset utilized in this study is separate from the enhancer data available through the VISTA Enhancer Browser, which is vital to mention. A sizable dataset with 43,011 enhancer candidates drawn from a wide range of samples was made available by the FANTOM5 group. These samples come from human participants and include 241 cell line samples, 432 primary cell samples, and 135 tissue-specific samples. Computationally, this hybrid CNN-DLSTM model's pseudocode is shown in the following outline.

```
# Import necessary libraries

import numpy as np
import tensorflow as tf

# Define the CNN-DLSTM model architecture

model = tf.keras.Sequential()

# Convolutional Layers

model.add(tf.keras.layers.Conv1D(filters=32, kernel_size=5, activation='relu', input_shape=(sequence_length, 4)))
model.add(tf.keras.layers.MaxPooling1D(pool_size=2))
model.add(tf.keras.layers.Conv1D(filters=64, kernel_size=3, activation='relu'))
model.add(tf.keras.layers.MaxPooling1D(pool_size=2))

# LSTM Layers

model.add(tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(64, return_sequences=True)))
model.add(tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(64, return_sequences=True)))

# Fully Connected Layers

model.add(tf.keras.layers.Flatten())
model.add(tf.keras.layers.Dense(128, activation='relu'))
model.add(tf.keras.layers.Dropout(0.5))
model.add(tf.keras.layers.Dense(1, activation='sigmoid'))

# Compile the model

model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

# Load and preprocess your DNA enhancer data

X_train, y_train = load_and_preprocess_data(train_data_path)
X_test, y_test = load_and_preprocess_data(test_data_path)

Train the model

model.fit(X_train, y_train, epochs=epochs, batch_size=batch_size, validation_split=0.2)

Evaluate the model on test data

loss, accuracy = model.evaluate(X_test, y_test)
print(f'Test loss: {loss}')
print(f'Test accuracy: {accuracy}')
```

The data was analyzed through Mathew's correlation coefficient principle using steps like precision recall, specificity, and accuracy. This method is commonly used to evaluate categorization accuracy. Binary classification labels "TP" & "TN" ["true positive" & "true negative"] and "FP" & "FN" ["false positive" & "false negative"] were generously used for measuring categorization accuracy (**Perreault Jr *et al.***, 2021). The observed numerical value ranges from negative to positive. The system outputs a numerical number from -1 to +1 based on input. We used 5-fold cross-validation to evaluate the model due to the dataset's restrictions and short datasets, which predicts outcomes by averaging five forecasts. The area under the curve (AUC), also known as the Receiver Operating Characteristics curve (ROC curve), was used to evaluate model performance. A higher AUC indicates model accuracy and quality. In the end, the average of five 5-fold cross-validated forecasts was calculated to yield the final prediction.

4. Theoretical Framework

The study proposed a Liver Cancer Prediction Framework. This framework is based on the premise that human DNA dataset does not readily reveal the existence of enhancers, which are DNA regions that bind to transcription factors. Within its particular AI-analyzed biological region, the enhancer sequence is in charge of boosting the level of transcriptional activity. The interaction of the enhancer sequence with cell transcription factors (TFs), which are linked to open chromatin areas, results in this increase. Sequence-level factors also have an impact on the enhancer sequence's efficacy. As they interact spatially with other regulatory elements, like promoters, in a three-dimensional environment, enhancers can exist in three different states: dormant, prepared, or active. For the recruitment of transcription factors (TFs) to begin the transcription of the target gene, the regulatory elements act as a stable framework (**Alharbi; Vakanski**, 2023). Hence, enhancers play a crucial role in the intricate gene expression program that takes place during human development. The intricate coding structures, their occurrence in other genetic elements, and the absence of a distinct enhancer sequence code, among several other factors, provide challenges for researchers in their endeavor to ascertain enhancer elements inside genes (**Pradhan *et al.***, 2023). Furthermore, the precise assessment of enhancer functioning and the precise determination of enhancer sites throughout the whole human AI-processed DNA dataset is of paramount importance.

The recommended model's robustness in this framework is enhanced by the use of data out of two distinct datasets, namely VISTA as well as FANTOM5 Enhancer Browser. These two strategies being investigated operates in two distinct stages. To capture the relevant attributes, the process of feature representation is conducted at the first stage on attributes such as k-mers, one-hot encoding, and shape. The second stage of the process employs hybrid CNN-Deep Long-Short Term Memory (CNN-DLSTM) DL model to forecast enhancers and their respective strengths. This forecast is produced based on the features emanated from the preceding round. When CNN is combined with LSTM networks, this new hybrid architecture called a CNN-DLSTM is created. Models with long short-term memory (LSTM) are optimized for processing time series and other forms of sequential data. In contrast, CNNs are frequently used for analyzing data's geographic properties. The RNN architecture was modified to create the long short-term memory (LSTM) neural network. RNNs are trained by feeding information into the network in a linear fashion until it reaches the output neurons. Calculated mistakes are then relayed backward to the network to change its settings. Information loops are included in the hidden levels of this particular network architecture. Loops allow information to flow in both directions, enabling the disguised state to collect historical data at a specific time step. Therefore, all results are dependent on previous, proven forecasts.

However, there are some limitations to how successfully RNNs can link beyond a given number of sequential steps. Since it explains the gradual loss of knowledge from earlier steps, gradient vanishing serves as the key explanation for the capacity of prediction models to capture short-term dependencies over time. As the depth of an RNN with activation functions grows, the loss function's gradient tends to approach zero. Long short-term memory neural networks (LSTM-NNs) are useful for fostering long-term associations. Since the lengthy Short-Term Memory (LSTM) model incorporates a memory unit and gate mechanism, it may efficiently record sequences with lengthy dependencies. Therefore, Long Short-Term Memory Neural Networks (LSTM-NNs) may learn from a large number of consecutive steps and selectively store or discard information using three gates and cell states. The integration of spatial and temporal components of data might yield valuable insights. Ultimately, the model undergoes training utilizing the provided spatial and temporal data.

Table 2: Tuned Hyperparameters for CNN.

Model Configuration	Output Shape
Input DNA sample	[196,100]
Convolutional Layer (8x7x1)	[196,8]
Convolutional Layer (16x11x1)	[196,16]
Dropout (rate = 0.5)	(1470)
Dense Layer (1)	1
Softmax (1)	1

Table 2 shows tuned hyperparameters for CNN suggesting that local features can be detected by convolutionizing input DNA sample with a kernel. Multiple kernels can be merged to make a filter. A CNN has several layers, but the convolutional and max-pooling layers are crucial. The convolutional layer uses a weight matrix and feature maps to extract and categorize important feature data. Spatial invariance can be achieved by using a maximum pooling layer with reduced feature map resolution. The neural network uses multiple fully connected layers after convolutional and max-pooling layers to better understand the dataset's non-linear combinations of high-level properties. Following DeepBind, this study built a three-layer CNN containing a convolutional layer, a ReLU activation as well as max-pooling layers. These last layers infer properties obtained as a result of input. CNN model input is the combination of results from different features and their representation-based methodologies.

The results section discusses in detail the schematic representation of this investigated analysis framework. It is shown that datasets utilized in this investigation included experimentally validated human enhancers from a wide array, which were obtained from the VISTA Enhancer Browser collection in 1747. The atlas of active enhancers from FANTOM5 was the source of the other used in this investigation. The other mentioned methods in this five-step rules for building a reliable benchmark dataset were used to select this dataset, which guarantees accurate training and testing of the suggested model.

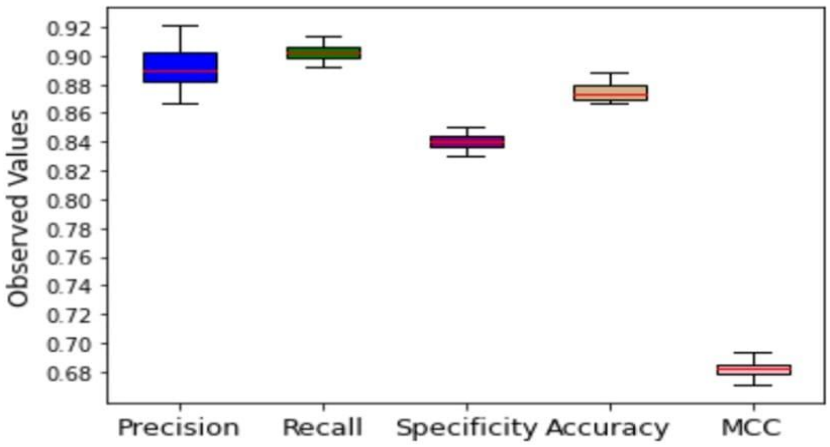
5. Results and Findings

The feature representation learning strategy incorporates elements from Word2Vector (W2V), One-Hot Encoding (**Ezziane**, 2006), and AI-driven shape analysis methods. In Table 3, the model is applied to data with varied k-mer lengths from 1 to 3 to 5 to 6 mers to discover the optimal k. The optimal k value may then be found. Grid-search revealed that the 3-mer structure performs better than all other k values (**Khalsan *et al.***, 2022).

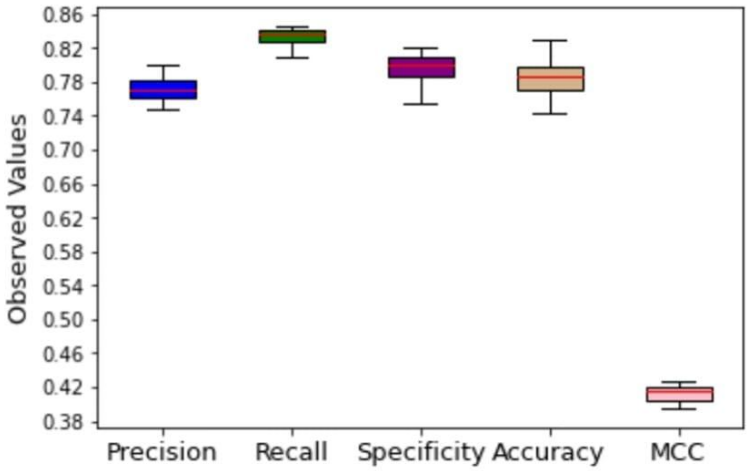
Table 3: Performance of the Suggested Model on Different Values of k-mer on the First and Second Stages.

Stage	k-mer	Precision	Specificity	Recall	Accuracy	MCC
First	1-mer	0.81	0.79	0.89	0.82	0.56
	3-mer	0.85	0.83	0.92	0.84	0.59
	5-mer	0.80	0.78	0.86	0.81	0.54
	8-mer	0.75	0.74	0.81	0.77	0.49
Second	1-mer	0.76	0.74	0.82	0.77	0.51
	3-mer	0.83	0.80	0.88	0.81	0.55
	5-mer	0.79	0.75	0.84	0.78	0.53
	8-mer	0.72	0.69	0.76	0.71	0.45

Five-fold cross-validation tested the model's 3-mer prediction ability. The dataset was divided into five equal sections, with one for testing and four for training. Figure 2 shows boxplots of 5-fold cross-validation findings. The ROC investigation reveals that Word2Vector encoding classifies data better than One-Hot Encoding or AI-driven shape analysis techniques (Figure 3). The hybrid Convolutional Neural Network (CNN) and Deep Long Short-Term Memory (DLSTM) model's first and second phases provide AUC values of 0.86 and 0.81 (Figure 3). This is achieved by combining multiple feature representation techniques from deep learning-based approaches. Therefore, the model that included the results of many methods outperformed one that relied just on any one of them. Integrating multiple AI-based encoding techniques significantly enhances model performance compared to using a single method. The resulting dataset in this study was utilized in the subsequent prediction stage once the IDs have been trained and tested. Utilizing an unbalanced dataset, the optimized hybrid CNN-DLSTM model under examination was improved for enhanced information extraction and classification accuracy.

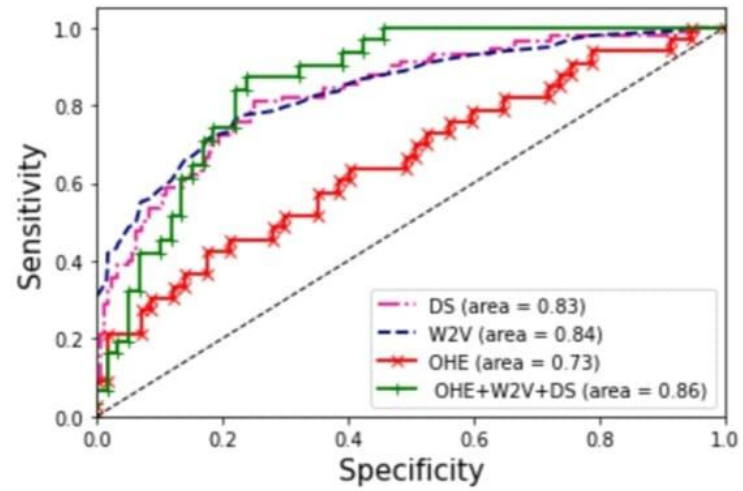


(a)

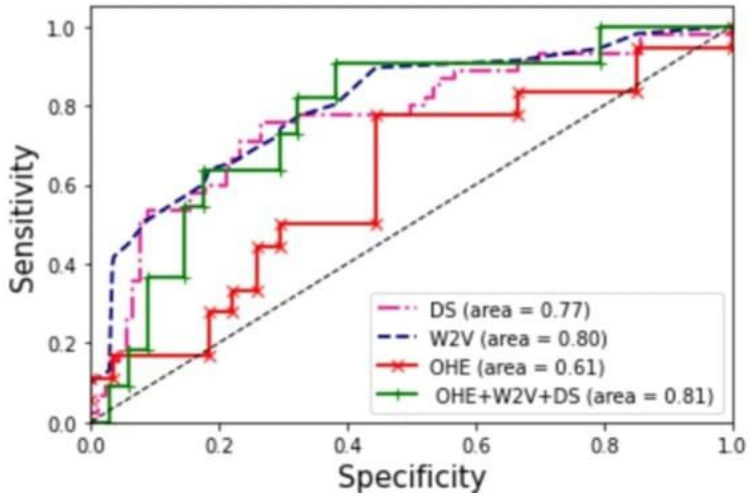


(b)

Figure 2: Five-cross Validation using 3-mer Data (a) and the Suggested Model's Boxplot (b).



(a)



(b)

Figure 3: ROC Curves for Word2vector, DNA Shape for (a) 1st Stage (b) 2nd Stage, One Hot Encoding and its Combination.

Taken from the FANTOM5, therefore we assessed the model on a separate balanced dataset to validate its applicability in a variety of contexts. The collection includes 1747 sequences with visible enhancer activity that were obtained from VISTA. Performance of the hybrid CNN-DLSTM and LSTM-NN techniques on independent datasets is shown in Table 4.

Table 4: Comparing the Effectiveness of the iEnhancer-2L Technique and the Hybrid CNN-DLSTM on a Test Dataset.

Stage	Method	Specificity	Accuracy	MCC
First	iEnhancer-2L [160]	0.80	0.76	0.59
	Hybrid CNN-DLSTM (Proposed Method)	0.83	0.82	0.66
Second	iEnhancer-2L [160]	0.75	0.74	0.31
	Hybrid CNN-DLSTM (Proposed Method)	0.80	0.80	0.42

Table 5: Comparison of Hybrid CNN-DLSTM with Different Machine Learning-based Classifiers.

Stage	Method	Specificity	Accuracy	MCC
First	Support Vector Machine	0.83	0.76	0.56
	KNN	0.90	0.75	0.51
	Ensembles for Boosting	0.74	0.73	0.48
	Random Forest	0.77	0.74	0.49
	Hybrid CNN-DLSTM (Proposed Method)	0.87	0.90	0.70
Second	Support Vector Machine	0.64	0.67	0.35
	KNN	0.63	0.67	0.31
	Ensembles for Boosting	0.60	0.64	0.30
	Random Forest	0.58	0.62	0.25

In order to examine alternative classifiers, classification methods like k-nearest neighbor (KNN), random forest, support vector machine (SVM), and ensembles for boosting are generally used in computational biology. These techniques are well-known for working and are frequently used in the industry. We investigated our model using several machine learning based classification techniques (**Ahmed *et al.***, 2016) to judge the efficiency of the suggested hybrid CNN-

DLSTM technique in recognizing enhancers as well as its strengths. The K-Nearest Neighbours (KNN) algorithm's efficiency depends on choosing the number of closest neighbors, whereas the Random Forest algorithm's success depends on the number of trees. We have chosen the ideal settings for each method so that we may compare them. Table 5 compares the performance of our model compared to other classifiers concerning accuracy, specificity, as well as Matthews correlation coefficient (**Perreault Jr *et al.***, 2021), illustrating its superiority. However, a hybrid CNN-LSTM model is thought to be preferable to using an LSTM unit alone. Our method also has the advantage of feeding information into the DL model by combining three feature representation strategies, that of specificity, accuracy, and MCC, as shown in Table 6 along with their values, though similar to other methods as seen in Table 5

Table 6 Hybrid CNN-DLSTM (Proposed Method).

Stage	Specificity	Accuracy	MCC
	0.80	0.80	0.42

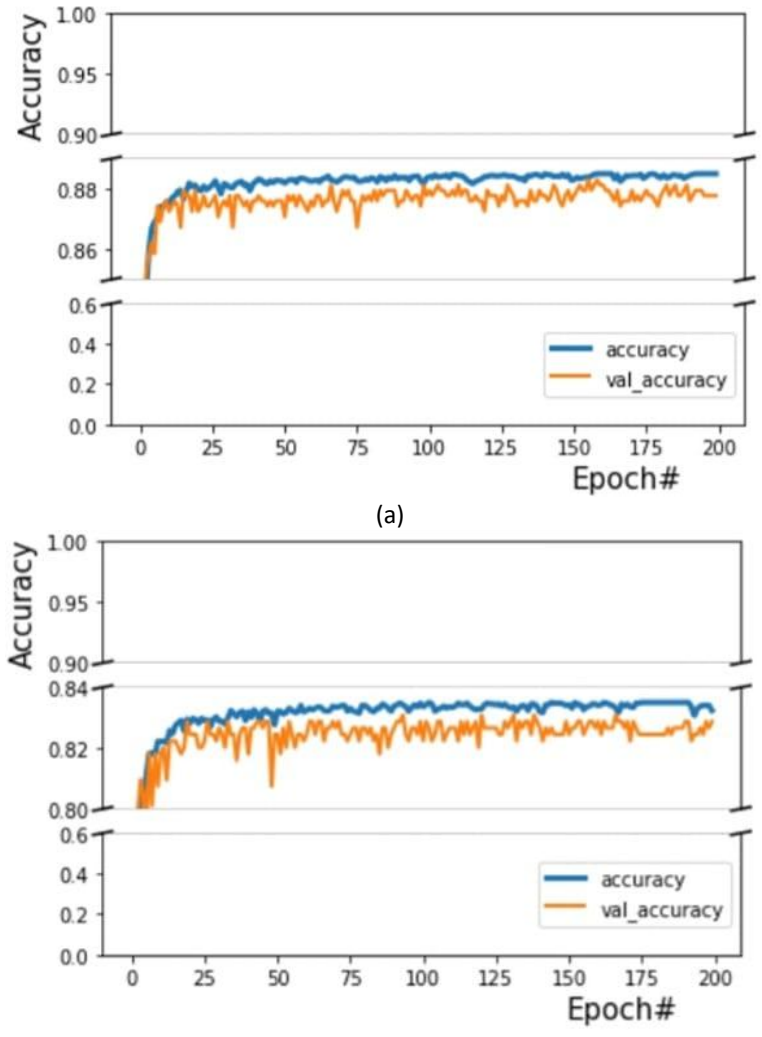


Figure 4. Plots showing accuracy for the proposed model's (a) first and (b) second stages, respectively

Figure 4 displays accuracy plots for the proposed model's first and second stages, respectively. When compared to SENIES, the suggested model behaves better in the first and second stages when it comes to precision, sensitivity, and MCC, respectively, according to the recommended model. The hybrid CNN-DLSTM model possesses inherent advantages in effectively and comprehensively capturing and learning the underlying properties. Potential future paths of this research encompass the development of an open-source web server to facilitate the utilization of the proposed hybrid CNN-DLSTM model. This endeavor has promise to enhance computational biology capabilities within the field of medical science.

6. Discussion

The study developed a DL-based system to facilitate the extraction of features utilizing three unique feature representation methodologies, namely CNN and DLSTM as shown in Figure 5. Additionally, this system enables the prediction of enhancers and their potency. The proposed method uses information garnered both from FANTOM5 along

the VISTA enhancer browser, which results in a more accurate categorization of enhancer components. By employing a word embedding model for expressing characteristics and successfully implementing a single-dimensional CNN towards the most advanced stage of classification, researchers have conducted more studies into the possible outcomes of the suggested technique. Our method finds its primary use in the fields of medicine, pharmaceutical research, and computational biology. This is due to the essential role enhancers play as cis-regulatory elements in the regulation of several genes.

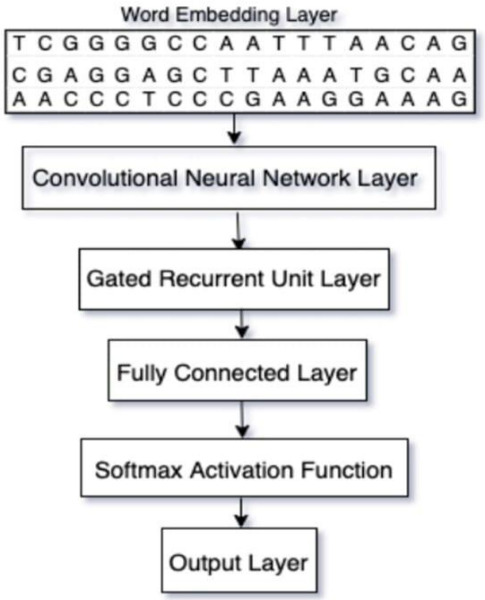
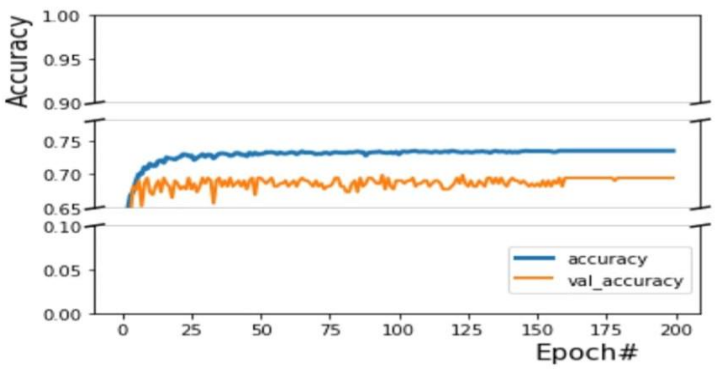
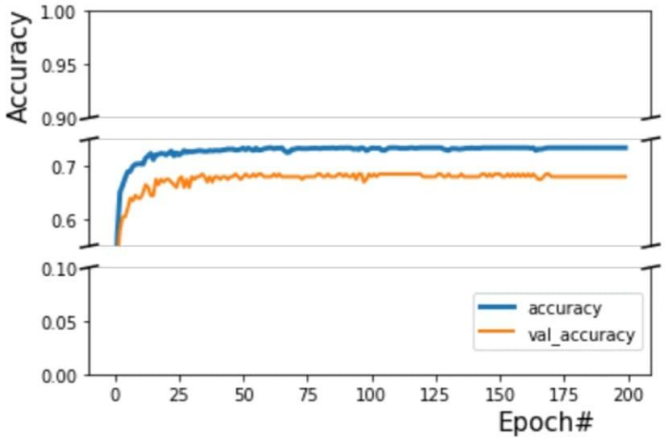


Figure 5: DL-based CNN-GRU model utilized using CNN-DLSTM.



(a)



(b)

Figure 6: Accuracy Plots for (a) First Stage of GRU based Model (c) Second Stage of GRU based Model.

Figure 6 shows an improvement to the training strategy. The present DL approach was facing an issue of over-fitting

due to the restricted amount of enhancer samples in a particular cell. The benchmark dataset was created using the method described in this article, which makes substantial use of FANTOM5 and VISTA enhancer browser data. After that, the model was trained using the proper hyper-parameters. Evidence of the model's advantage over conventional training techniques comes from the evaluation of the model on an independent dataset. The combination of CNNs and Deep Long Short-Term Memory (DLSTM) networks is a recent development in the field of architecture. Table 6 shows the evidence that that DLSTM, a kind of recurrent neural network, improves the model's performance by successfully capturing long-range sequence interactions.

Table 6: Comparison of Hybrid CNN-DLSTM with LSTM based Model for Enhancer Identification.

Stage	Method	Specificity	Accuracy	MCC
First	iEnhancer-EBLSTM [225]	0.80	0.76	0.533
	Hybrid CNN-DLSTM (Proposed Method)	0.89	0.90	0.70
Second	iEnhancer-EBLSTM [225]	0.54	0.66	0.31
	Hybrid CNN-DLSTM (Proposed Method)	0.80	0.82	0.42

Table 7 and Table 8 compare Hybrid CNN-DLSTM with SENIES for first and second stages, suggesting that latter can be utilized as a proposed computational method to identify enhancer regions within DNA sequences by leveraging information about the 3D structure of the DNA, replacing the traditional sequence data, in order to achieve better prediction of both the presence of enhancers and their strength.

Table 7: Comparison of Hybrid CNN-DLSTM with SENIES for First Stage.

Method	Feature Representation	AUC	Accuracy	MCC
SENIES [212]	One hot encoding (**Ezziane**, 2006)	0.80	73.71	0.48
	k-mer	0.82	75.81	0.52
	DNAshape	0.81	73.56	0.48
	All	0.82	76.81	0.54
Hybrid	One hot encoding (**Ezziane**, 2006)	0.72	75.69	0.51
CNN-DLSTM	k-mer	0.83	87.49	0.67
	DNAshape	0.82	81.57	0.58
	All	0.85	89.71	0.695

Table 8: Comparison of Hybrid CNN-DLSTM with SENIES for second stage.

Method	Feature Representation	AUC	Accuracy	MCC
SENIES	One hot encoding	0.64	57.96	0.164
	k-mer	0.67	62.51	0.254
Hybrid	DNAshape	0.68	62.04	0.244
CNN-DLSTM	All	0.74	68.24	0.381
	One hot encoding	0.60	63.48	0.306
	k-mer	0.81	79.11	0.411
	DNAshape	0.76	76.37	0.398
	All	0.82	81.41	0.42

For the purpose of liver cancer detection, algorithmic data analysis was used for the initial identification of predictive elements, and it was then used to validate the predictions and make the research more easily applicable. Our research efforts focused on the study of cancer, with special attention to the field of AI-driven radiology and medical imaging. In the developing subject of AI-powered radiology models, cancer-related predictive data and imaging data are combined to improve forecast accuracy and precision Based on the importance of exploration in localization and the degree of illness propagation, this work presents an algorithmic approach for classifying the slices into six different groups. The procedure of practicing a deep neural network (DNN) model utilizing a dataset made up of CT scans of liver cancer achieves the aforementioned goal.

The TCIA dataset, which includes volume images (140 CT) of patients with probable liver cancer, was used in this investigation. A total of 140 images makes up the collection, with 131 of them being computed tomography (CT) images and the remaining 9 being positron emission tomography-computed tomography (PET-CT) scans. Anatomical structures such as the lung, bone, liver, kidney, bladder, and brain have been labeled on the matching segmentation images that go along with these images. Additionally, there is a lot of variation in the dataset because it was collected from different CT machines and AI-based imaging facilities. The development of a reliable model for learning purposes is significantly hampered by this variability. The dataset is made up of volumetric images with various abdominal and whole-body slices. Patients without at least a single exam of an organ containing a lesion, benign or malignant, were excluded from the dataset.

The NIFTI-1-compliant 32-bit floating-point data was used to create the pictures in this collection (See Figure 7 and Figure 8). To analyze CT scans in the NIFTI-1 format, radiologists all over the world frequently use DICOM software. Each gantry rotation produces several slices that make up a computed tomography (CT) volume. The radiologist's expertise and amount of training will determine how well they can extract useful information from a group of slices. The computerized tomography (CT) imager generates several types of aberrations and noise that enter CT photos

throughout the image gathering process, reducing the effectiveness of a particular ML model (**Lee; Jang**, 2022). Before training the network, pre-processing methods like histogram equalization and Gaussian filtering were used on the volume picture data to reduce the influence of these errors. The dataset was acquired using different CT scanner settings, resulting in volume images with different contrast levels. On the volume data (**Tran *et al.***, 2023), histogram equalization was used to change the contrast.

In this investigation, the original image's 512x512 pixels were downsized to 32x32 pixels before being used in the DL model. CNN models can be evaluated for performance using a preprocessing method that involves scaling the initial image toward the neural network. The ability to slice and segment photos more efficiently computationally thanks to the retention of image contents despite scaling. A volume slice as well as its segmented picture are depicted together in Figure 7 to Figure 9. The liver, kidney, and bone are the organs that are visible in these particular scans. The bladder, lung, and brain are invisible inside the scan, though patients with liver cancer are subjected to full-body and abdominal imaging. While imaging the liver and other organs that may be vulnerable to the development of tumors is the primary goal of using a CT scanner, it is important to keep in mind that abdominal and whole-body scans also include imaging of several other organs, including the lung, bladder, kidney, bone, and brain. Although liver cancer has the potential to migrate to the lungs and then to the bones, the risk of it doing so is quite low. Consequently, a performance examination of the liver, lung, and bone has all been done.

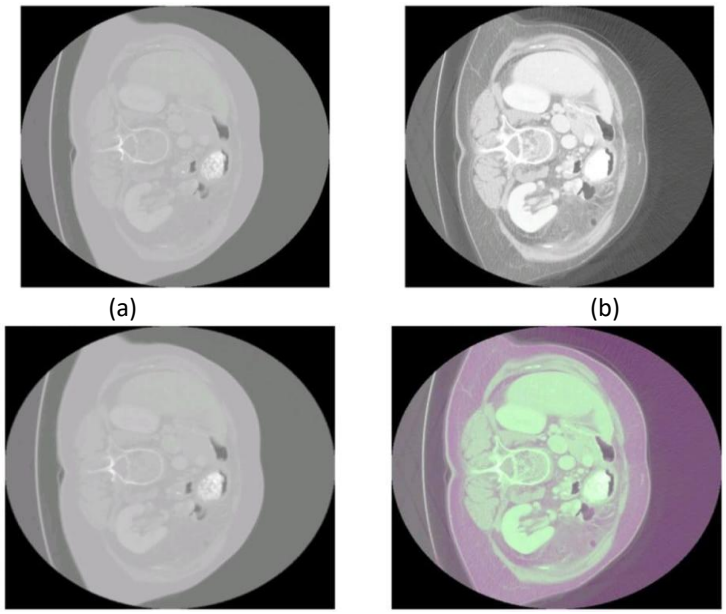


Figure 7: Volume Slices of (a) Original (b) Histogram Equalized. (c) Gaussian filtered (d) Data augmented.

It is not required for radiologists to give each of the 63,503 volume slices the same amount of attention. Figure 7(a) shows a cross-sectional view of the liver tissue, whereas Figure 7(c) depicts a cross-section showing organs devoid of the liver. Slices of Figures 7(a) and 7(c), respectively, are shown in multi-color histograms with bars of various colors in Figures 7(b) and 7(d). The organ colors from the relevant segmented pictures were used to color the horizontal lines within the histogram for consistency and clarity. The histogram shown in Figure 8 reveals the (a) Original (b) Segmented proportions of various organs distributed over the dataset's 63503 volume images. It demonstrates that the liver is visible in about 30% of the dataset's slices, while other organs are shown in 70% of the images.

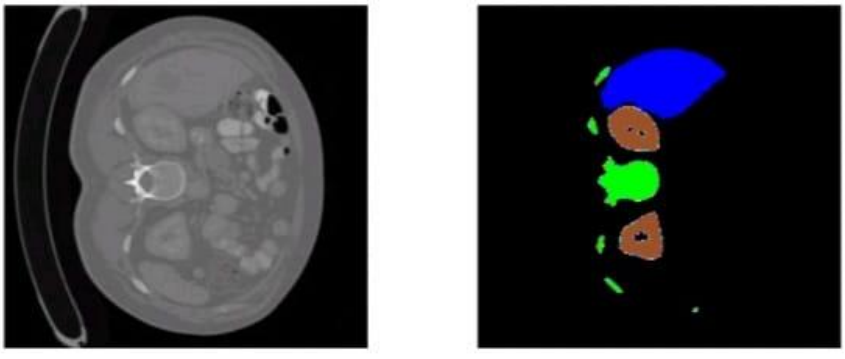


Figure 8: (a) Original (b) Segmented.

Slice sorting procedure significance is indicated by the ratio of slices with liver to those without liver. This reduces the amount of time needed for the radiation oncologist to diagnose liver cancer and for radiation therapy. The colors in the segmented slice pictures of Figures 8(a) and 8 (b) are coherent with the colors in the bars in Figure 9 (b) and Figure 9 (d) once more.

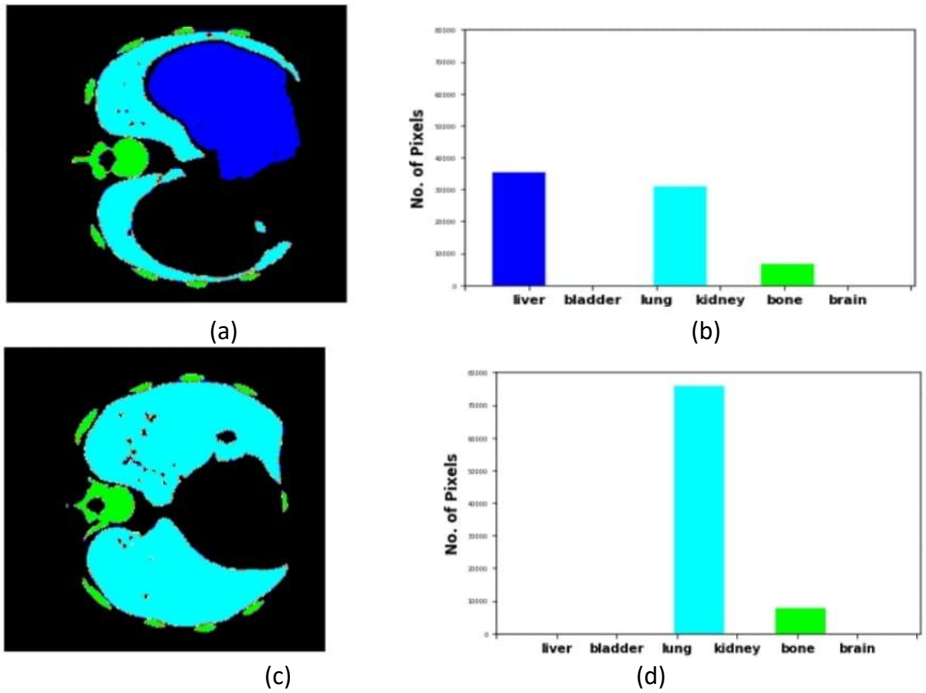


Figure 9: (a) Slice with Liver (b) Slice Histogram (c) Slice without Liver (d) Slice Histogram.

Figure 9 confirm the resilience of the proposed model in categorising the test image, the accuracy is determined for both training and validation dataset of original volume image, histogram equalized volume image, filtered image and data augmented image.

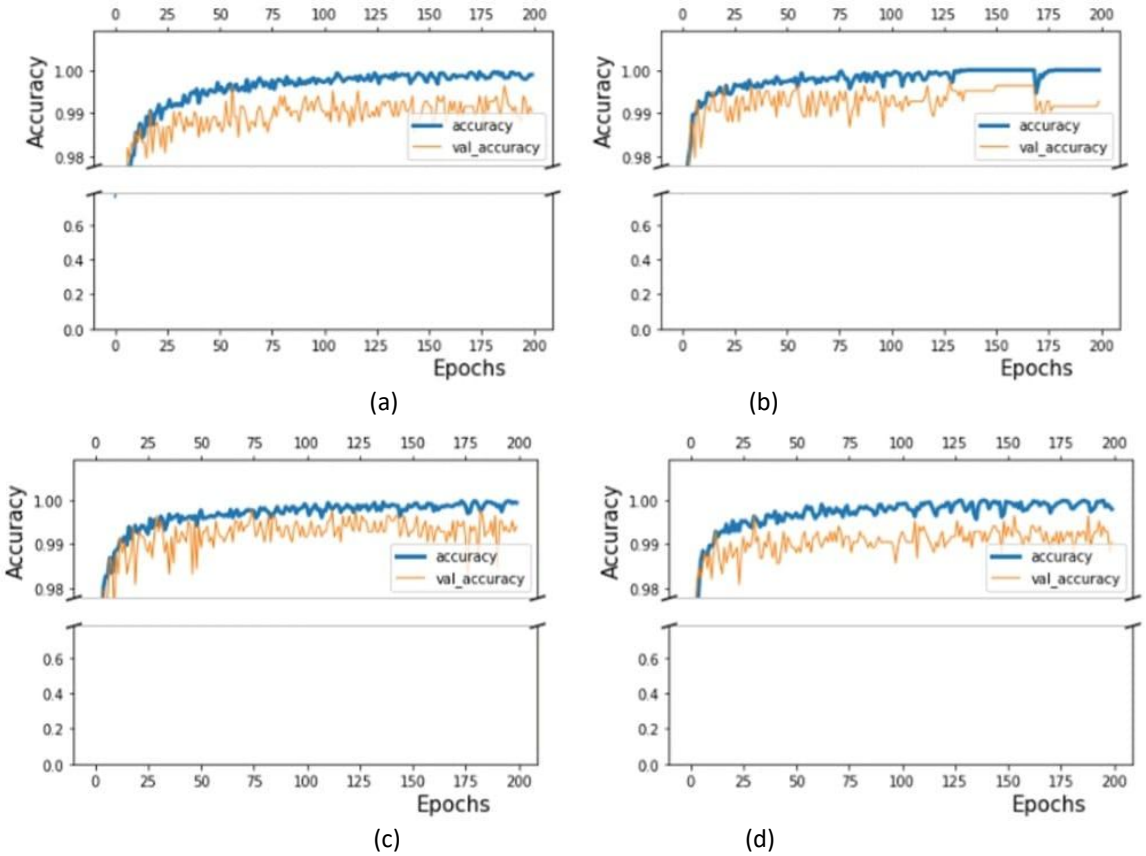


Figure 10: Comparison of Model Accuracy on (a) Original volume Slices (b) Histogram Equalized Volume Slices (c) Gaussian Filtered Slices (d) Volume Slices with Data Augmentation.

Figure 10 depicts receiver operating characteristics (ROC), which can be readily observed in Figure 10(d), having optimal cut-off resulting into maximum rate of True Positives with lowest False Positives. The classification results in terms of accuracy, precision, sensitivity, specificity, true positive rate, false-negative rate, and F1 score have been computed using confusion matrix for each category of volume slice. For the localization of liver cancer and its spread to other sections of the body, the organs which are predominantly impacted are liver, lung and bone. Therefore, the confusion matrix is computed for the existence of three organs viz. liver, lung, and bone.

Table 9: Experimentally Derived Validation Accuracy and Model Loss.

Metrics	Volume Slices of Original Image	Volume Slices of Histogram Equalized Image	Volume Slices of Filtered Image	Data Augmented Volume Slices
Accuracy	0.99	0.99	0.99	0.99
Loss	0.04	0.04	0.04	0.02

The computed metrics are provided in Table 9 for the original volume slices, histogram-equalized volume slices, filtered volume slices, and data supplemented volume slices, respectively. For each category of volume segment, the values of true positives (TP), true negatives (TN), false positives (FP), and false negatives (**AlAfnan; MohdZuki**, 2024) were calculated. Utilizing TP, TN, FP, and FN, the metrics of precision, recall, specificity, sensitivity, accuracy, and F1 measure are assessed.

Table 10: Metrics for Various Values of Variance Over Data Augmented Volume Slices.

Metric	Variance = 5	Variance = 10	Variance = 15
	Liver	Lung	Bone
True Positive	175	82	341
True Negative	485	617	272
False Positive	15	19	10
False Negative	55	12	17
Precision	0.93	0.82	0.89
Recall (Sensitivity)	0.75	0.89	0.95
Specificity	0.03	0.16	0.026
Accuracy	0.91	0.90	0.92
F-measure	0.84	0.85	0.91

Table 10 demonstrates metrics for various values of variance over data augmented volume slices.

While the model is trained using the additional volume slices from the collected information, the results show a much higher validation performance of 99.1% contrasted to the efficiency of 98.7% achieved while training using the original volume slices. Data augmented volume slice dataset has a test accuracy of 93.1% overall, which is higher than other volume slices.

7. Conclusion

To solve the problem of identifying patterns in AI-driven data processing, researchers have turned to deep learning (DL) techniques. Various AI-based repositories and algorithmic pattern detection systems provide datasets used by these techniques. As a result of the high degree of dependency between algorithmic sequences, CNN has proven to be inadequate as a predictor of complex patterns. The use of DL-based frameworks with CNN and deep long short-term memory (DLSTM) outperformed state-of-the-art approaches on a variety of evaluation metrics. These metrics included precision, recall, specificity, accuracy, and Matthew's correlation coefficient. Moreover, the significance of the given metrics is clearer when they are compared against each other to evaluate performance outlook. Therefore, it is recommended that the incorporation of some of the additional AI-driven methods will provide better results with more accurate predictions. Future research should focus on integrating predictive models for real-time data analysis, which would help improve the development and effectiveness of such models. Radiation oncologists can more precisely target tumors by using an AI-based automated system to classify 3D CT scans of the liver that show signs of malignancy. The use of this technology aids radiologists in making correct diagnoses. Recent advancements in AI and machine learning make both of these uses possible.

However, the classification's effectiveness in lowering the rate of false negatives will depend on how well it functions in actual use. This paper presents a model-based, automated method for classifying multiple organs in liver cancer CT imaging. This method gives the model the ability to identify which parts of the data are useful and which aren't. Oncologists' ability to zero in on a specific subset of a large dataset improves when only a portion of that dataset is presented to them. That makes it possible for people to quickly and easily move between many different areas. The network is trained using multi-data that includes liver cancer-related CT scans taken in three dimensions. Multiple AI-driven CT scanners were used to collect the pictures, each with its own unique medical parameters and machine-specific configurations. The effectiveness of the proposed method is put through its paces with a variety of test images. The results of a performed study using a wide range of different datasets have shown that the proposed method has significant advantages for drastically cutting down on false negatives, lowering the rate of unnecessary medical

interventions, and saving time, effort, and money. In some cases, it might also prevent unnecessary biopsies from being performed. The results of this study are anticipated to aid radiologists and oncologists in the early diagnosis and treatment of liver cancer.

7.1. Future Scope and Implications

Potential future paths of this research encompass the development of an open-source web server to facilitate the utilization of the proposed hybrid CNN-DLSTM model. This method has promise for enhancing AI-driven computational analysis capabilities in information processing and predictive modeling. By using a real-time AI-assisted adaptive system, it is possible to modify the proposed model to take dynamic variations into account and reduce errors in automated decision-making. Principal Component Analysis (PCA) has the potential to be employed to strengthen the strategy on the supplied data by optimizing feature selection and dimensionality reduction. The dataset used here is also somewhat small; in the future, the model will continue to be trained and tested on larger AI-driven datasets specific to various predictive tasks. However, there is much potential for future research to identify pattern propagation across different data categories, as the present investigation just examines specific classification cases. Because certain data structures tend to be sparse, the proposed technique faces challenges in accurately classifying them. The neural network's classification results can be improved by adding more convolutional layers and refining hyperparameter tuning for better feature extraction. Additionally, this research is currently limited to specific types of input data, while it can be easily extended to multi-modal AI models for broader classification applications. Recognizing the practical significance and impact of AI-learned features is one way this research can be expanded further, providing valuable insights into automated data analysis and deep learning applications.

References

Adadi, Amina. (2021). "A survey on data-efficient algorithms in big data era". *Journal of Big Data,* v. 8, n. 1, pp. 24. *https://doi.org/10.1186/s40537-021-00419-9*

Ahmed, Umair; Zin, Md Lazim Mohd; Majid, Abdul Halim Abdul. (2016). "Impact of Intention and Technology Awareness on Transport Industry's E-Service: Evidence From an Emerging Economy". *The Journal of Industrial Distribution & Business,* v. 7, n. 3, pp. 13-18. *https://doi.org/10.13106/ijidb.2016.vol7.no3.13*

AlAfnan, Mohammad Awad; MohdZuki, Siti Fatimah. (2024). "Malaysia's National Blockchain Roadmap: A Critical Discourse Analysis of Focus, Goals, and Challenges". *World Journal of English Language,* v. 14, n. 5, pp. 482-492. *https://doi.org/10.5430/wjel.v14n5p482*

Alharbi, Fadi; Vakanski, Aleksandar. (2023). "Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review". *Bioengineering,* v. 10, n. 2, pp. 173. *https://doi.org/10.3390/bioengineering10020173*

Bakrania, Anita; Joshi, Narottam; Zhao, Xun; Zheng, Gang; Bhat, Mamatha. (2023). "Artificial intelligence in liver cancers: Decoding the impact of machine learning models in clinical diagnosis of primary liver cancers and liver cancer metastases". *Pharmacological Research,* v. 189, pp. 106706. *https://doi.org/10.1016/j.phrs.2023.106706*

Berente, Nicholas; Gu, Bin; Recker, Jan; Santhanam, Radhika. (2021). "Managing Artificial Intelligence". *MIS Quarterly,* v. 45, n. 3, pp. 1433-1450. *https://doi.org/10.25300/MISQ/2021/16274*

Brynjolfsson, Erik; McAfee, Andrew. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. WW Norton & Company. *https://wwnorton.com/books/the-second-machine-age*

Chen, Fulong; Tang, Yuqing; Wang, Canlin; Huang, Jing; Huang, Cheng; Xie, Dong; Wang, Taochun; Zhao, Chuanxin. (2021). "Medical Cyber–Physical Systems: A Solution to Smart Health and the State of the Art". *IEEE Transactions on Computational Social Systems,* v. 9, n. 5, pp. 1359-1386. *https://doi.org/10.1109/TCSS.2021.3122807*

Cheng, Binglin; Zhou, Peitao; Chen, Yuhan. (2022). "Machine-learning algorithms based on personalized pathways for a novel predictive model for the diagnosis of hepatocellular carcinoma". *BMC Bioinformatics,* v. 23, n. 1, pp. 248. *https://doi.org/10.1186/s12859-022-04805-9*

Das, Bihter; Toraman, Suat. (2022). "Deep transfer learning for automated liver cancer gene recognition using spectrogram images of digitized DNA sequences". *Biomedical Signal Processing and Control,* v. 72, pp. 103317. *https://doi.org/10.1016/j.bspc.2021.103317*

Deng, Yang; Qin, Hang-Yu; Zhou, Yan-Yan; Liu, Hong-Hong; Jiang, Yong; Liu, Jian-Ping; Bao, Ji. (2022). "Artificial intelligence applications in pathological diagnosis of gastric cancer". *Heliyon,* v. 8, n. 12, pp. e12431. *https://doi.org/10.1016/j.heliyon.2022.e12431*

Ding, Li; Wendl, Michael C.; Koboldt, Daniel C.; Mardis, Elaine R. (2010). "Analysis of next-generation genomic data in cancer: accomplishments and challenges". *Human Molecular Genetics,* v. 19, n. R2, pp. R188-R196. *https://doi.org/10.1093/hmg/ddq391*

Durge, Aditi R.; Shrimankar, Deepti D.; Sawarkar, Ankush D. (2022). "Heuristic Analysis of Genomic Sequence Processing Models for High Efficiency Prediction: A Statistical Perspective". *Current Genomics,* v. 23, n. 5, pp. 299-317. *https://doi.org/10.2174/1389202923666220927105311*

Ezziane, Zoheir. (2006). "Applications of artificial intelligence in bioinformatics: A review". *Expert Systems with Applications,* v. 30, n. 1, pp. 2-10. *https://doi.org/10.1016/j.eswa.2005.09.042*

Gao, Qing; Yang, Luyu; Lu, Mingjun; Jin, Renjing; Ye, Huan; Ma, Teng. (2023). "The artificial intelligence and machine learning in lung cancer immunotherapy". *Journal of Hematology & Oncology,* v. 16, n. 1, pp. 55. *https://doi.org/10.1186/s13045-023-01456-y*

Hanczar, Blaise; Bourgeais, Victoria; Zehraoui, Farida. (2022). "Assessment of deep learning and transfer learning for cancer prediction based on gene expression data". *BMC Bioinformatics,* v. 23, n. 1, pp. 262. *https://doi.org/10.1186/s12859-022-04807-7*

Higgins, David; Madai, Vince I. (2020). "From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare". *Advanced Intelligent Systems,* v. 2, n. 10, pp. 2000052. *https://doi.org/10.1002/aisy.202000052*

Ho, Y. C.; Pepyne, D. L. (2002). "Simple Explanation of the No-Free-Lunch Theorem and Its Implications". *Journal of Optimization Theory and Applications,* v. 115, n. 3, pp. 549-570. *https://doi.org/10.1023/A:1021251113462*

Hou, Guojie; Jia, Liye; Zhang, Yanan; Wu, Wei; Zhao, Lin; Zhao, Juanjuan; Wang, Long; Qiang, Yan. (2023). "Deep learning approach for predicting lymph node metastasis in non-small cell lung cancer by fusing image–gene data". *Engineering Applications of Artificial Intelligence,* v. 122, pp. 106140. *https://doi.org/10.1016/j.engappai.2023.106140*

Jöhnk, Jan; Weißert, Malte; Wyrtki, Katrin. (2021). "Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors". *Business & Information Systems Engineering,* v. 63, n. 1, pp. 5-20. *https://doi.org/10.1007/s12599-020-00676-7*

Khalsan, Mahmood; Machado, Lee R; Al-Shamery, Eman Salih; Ajit, Suraj; Anthony, Karen; Mu, Mu; Agyeman, Michael Opoku. (2022). "A survey of machine learning approaches applied to gene expression analysis for cancer prediction". *IEEE Access,* v. 10, pp. 27522-27534. *https://doi.org/10.1109/ACCESS.2022.3146312*

Lee, Sung Hak; Jang, Hyun-Jong. (2022). "Deep learning-based prediction of molecular cancer biomarkers from tissue slides: A new tool for precision oncology". *Clinical and molecular hepatology,* v. 28, n. 4, pp. 754-772. *https://doi.org/10.3350/cmh.2021.0394*

Lee, Young-Ji; Park, Jun-Hyung; Lee, Seung-Ho. (2022). "A Study on the Prediction of Cancer Using Whole-Genome Data and Deep Learning". *International Journal of Molecular Sciences,* v. 23, n. 18, pp. 10396. *https://doi.org/10.3390/ijms231810396*

Li, Rufeng; Li, Lixin; Xu, Yungang; Yang, Juan. (2021). "Machine learning meets omics: applications and perspectives". *Briefings in Bioinformatics,* v. 23, n. 1, pp. bbab460. *https://doi.org/10.1093/bib/bbab460*

Moslehi, Fateme; Haeri, Abdorrahman. (2020). "A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection". *Journal of Ambient Intelligence and Humanized Computing,* v. 11, n. 3, pp. 1105-1127. *https://doi.org/10.1007/s12652-019-01364-5*

Painuli, Deepak; Bhardwaj, Suyash; köse, Utku. (2022). "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review". *Computers in Biology and Medicine,* v. 146, pp. 105580. *https://doi.org/10.1016/j.compbiomed.2022.105580*

Perreault Jr, William D; Cannon, Joseph P; McCarthy, E Jerome. (2021). *Essentials of Marketing: A Marketing Strategy Planning Approach*. McGraw-Hill. *https://www.mheducation.com/highered/product/Essentials-of-Marketing-Cannon.html*

Pradhan, Kanchan; Chawla, Priyanka; Rawat, Sanyog. (2023). "A deep learning-based approach for detection of lung cancer using self adaptive sea lion optimization algorithm (SA-SLnO)". *Journal of Ambient Intelligence and Humanized Computing,* v. 14, n. 9, pp. 12933-12947. *https://doi.org/10.1007/s12652-022-04118-y*

Rajpurkar, Pranav; Chen, Emma; Banerjee, Oishi; Topol, Eric J. (2022). "AI in health and medicine". *Nature Medicine,* v. 28, n. 1, pp. 31-38. *https://doi.org/10.1038/s41591-021-01614-0*

Sapoval, Nicolae; Aghazadeh, Amirali; Nute, Michael G.; Antunes, Dinler A.; Balaji, Advait; Baraniuk, Richard; Barberan, C. J.; Dannenfelser, Ruth; Dun, Chen; Edrisi, Mohammadamin; Elworth, R. A. Leo; Kille, Bryce; Kyrillidis, Anastasios; Nakhleh, Luay; Wolfe, Cameron R.; Yan, Zhi; Yao, Vicky; Treangen, Todd J. (2022). "Current progress and open challenges for applying deep learning across the biosciences". *Nature Communications,* v. 13, n. 1, pp. 1728. *https://doi.org/10.1038/s41467-022-29268-7*

Sun, Tara Qian; Medaglia, Rony. (2019). "Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare". *Government Information Quarterly,* v. 36, n. 2, pp. 368-383. *https://doi.org/10.1016/j.giq.2018.09.008*

Tran, Thi-Oanh; Vo, Thanh Hoa; Le, Nguyen Quoc Khanh. (2023). "Omics-based deep learning approaches for lung cancer decision-making and therapeutics development". *Briefings in Functional Genomics,* v. 23, n. 3, pp. 181-192. *https://doi.org/10.1093/bfgp/elad031*

Velichko, Yuri S.; Gennaro, Nicolo; Karri, Meghana; Antalek, Matthew; Bagci, Ulas. (2023). "A Comprehensive Review of Deep Learning Approaches for Magnetic Resonance Imaging Liver Tumor Analysis". *Advances in Clinical Radiology,* v. 5, n. 1, pp. 1-15. *https://doi.org/10.1016/j.yacr.2023.06.001*

Wang, Liyang; Song, Danjun; Wang, Wentao; Li, Chengquan; Zhou, Yiming; Zheng, Jiaping; Rao, Shengxiang; Wang, Xiaoying; Shao, Guoliang; Cai, Jiabin; Yang, Shizhong; Dong, Jiahong. (2023). "Data-Driven Assisted Decision Making for Surgical Procedure of Hepatocellular Carcinoma Resection and Prognostic Prediction: Development and Validation of Machine Learning Models". *Cancers,* v. 15, n. 6, pp. 1784. *https://doi.org/10.3390/cancers15061784*

Yin, Qingyan; Chen, Wangwang; Zhang, Chunxia; Wei, Zhi. (2022). "A convolutional neural network model for survival prediction based on prognosis-related cascaded Wx feature selection". *Laboratory Investigation,* v. 102, n. 10, pp. 1064-1074. *https://doi.org/10.1038/s41374-022-00801-y*