

ENSDEEP-ICAM: ENSEMBLE DEEP LEARNING IDENTIFIER FOR IDENTIFICATION OF MUTATION TO DETECT COLORECTAL ADENOCARCINOMA PROGRESSION

Asghar Ali Shah¹, Muhammad Usman Hashmi², Farman Ali³, Adil Khan⁴,
Zar Nawab Khan Swati⁵

¹Department of Computer Science, Beaconhouse International College, Islamabad, Pakistan

^{2,*3,4}Department of Computer Science, Bahria University, Islamabad, Pakistan

⁵Department of Computer Science, Karakoram International University, Gilgit-Baltistan, Pakistan

³farman.buic@bahria.edu.pk, farman335@yahoo.com

DOI: <https://doi.org/10.5281/zenodo.23007710>

Keywords

Colorectal cancer, Feature Representation, Artificial Intelligence, Machine Learning, Deep Learning

Article History

Received: 19 June 2026

Accepted: 10 September 2026

Published: 28 September 2026

Copyright @Author

Corresponding Author: *
Farman Ali

Abstract

Colorectal cancer is the third highest widespread disease killed millions of human beings worldwide. Early detection of cancer can greatly improve survival rates. Cancer can be caused by changes in the genetic code within cells, which can be caused by mutations in the gene sequences. These mutations can be the result of errors during replication or recombination of the genome. Some mutations, referred to as cancer driver mutations, can result in cancer. Several computational methods have been devoted for identifying colorectal adenocarcinoma mutation. However, the existing approaches are inadequate to detect the mutated reigns accurately. In this study, we developed a novel ensemble deep learning-based predictor called EnsDeep-iCAM for targeting colorectal adenocarcinoma mutations. The proposed method use effective feature extraction algorithms namely position relative incidence matrix, statistical moment, reverse position relative incidence matrix, frequency vector, accumulative absolute position incidence vector, and reverse accumulative absolute position incidence vector to encode the biological mutated sequences. Then we trained the models by using different variants of deep learning algorithms such as long-short-term memory network (LSTM), bi-directional LSTM (BiLSTM), gated recurrent units (GRU) and the ensemble of all three models (Ens-Deep). BiLSTM outperformed the implemented deep learning and machine learning classifiers using ten-fold cross-validation (10-FCVT), independent test and self-consistency testing (SCT) respectively. The proposed model achieved an accuracy and MCC of 96% and 0.91 on training dataset and 98% and 0.97 on testing dataset respectively. We believe that EnsDeep-iCAM protocol will provide valuable insights in characterization of large-scale unannotated mutated genes in particular and pharmaceutical drug discovery in general.

1 Introduction

Machine learning is a powerful tool that can be used to address a variety of crucial concerns in different disciplines. A kind of cancer called adenocarcinoma begins in secretory cells. The third most prevalent type of cancer in people is colorectal adenocarcinoma (CRC), which

develops when cells in the large intestine grow out of control. The incidence of colorectal cancer notably rises after the age of 50, with the average age at diagnosis being approximately 70 years in developed regions. According to data from 2012, Oceania, North America, and Europe exhibited the highest age-adjusted incidence rates of

colorectal cancer, reaching 41.0 per 100,000 for men and 29.2 per 100,000 for women. Conversely, regions such as south-central Asia and Western Africa reported the lowest rates, with incidence rates at 4.5 per 100,000 for males and 3.8 per 100,000 for females. Colorectal cancer manifests

when cells in the large intestine undergo uncontrolled proliferation. In contrast, regions such as south-central Asia and western Africa had notably lower rates, with 4.5 per 100,000 for males and 3.8 per 100,000 for females. (as shown in Figure 1) [2].

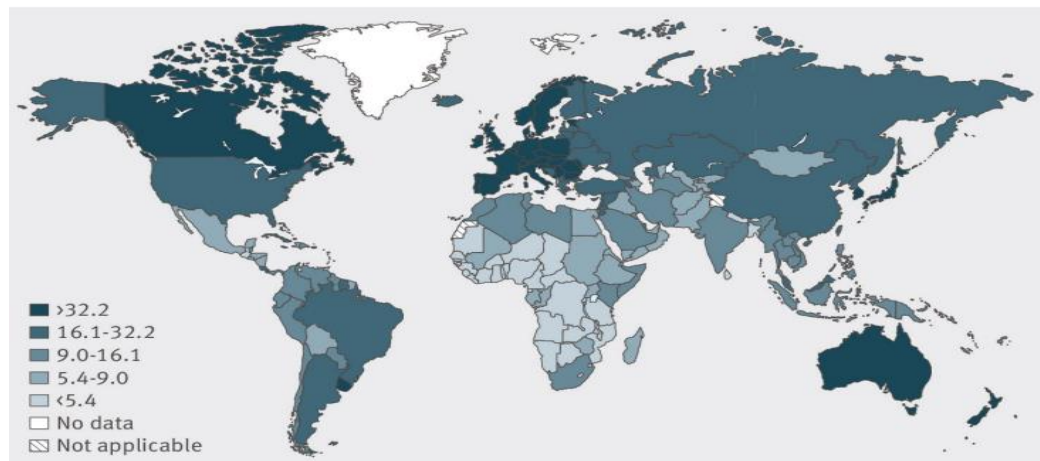


Figure 1: THE estimated global age-standardized incidence of CRC in males in 2012, expressed as the number of cases per 100,000 populations, is shown (adapted from GLOBOCAN 20121).

Regions that previously had low rates of colorectal cancer, the frequency of colorectal cancer has rapidly increased in areas that historically had low rates of the illness, including Spain and various nations in Eastern Europe and East Asia. This has been associated with consuming high-fat diets abundant in red and processed meats, leading a sedentary lifestyle, excessive alcohol consumption, and smoking. In contrast, illness incidence has plateaued or even decreased in the US and other industrialized nations, which may be related to greater use of diagnostic endoscopy and polypectomy[3] [4].

Colonoscopy is typically the preferred method for imaging in patients suspected of having colon

cancer and for obtaining biopsies of any suspicious lesions. A second colonoscopy, CT colonography, or a barium enema could be suggested to patients who had an unsatisfactory colonoscopy as a substitute[5].

Artificial intelligence (AI) has the potential to revolutionize the field of medicine, aimed at detecting various illnesses within the human body, numerous AI approaches have been created. This work employs machine learning to find CRC at developing state. The human gene consists of sequences of nucleotides, and any changes in these sequences are called mutations, which can often lead to cancer[6]. Figure 2 demonstrates mutation process.

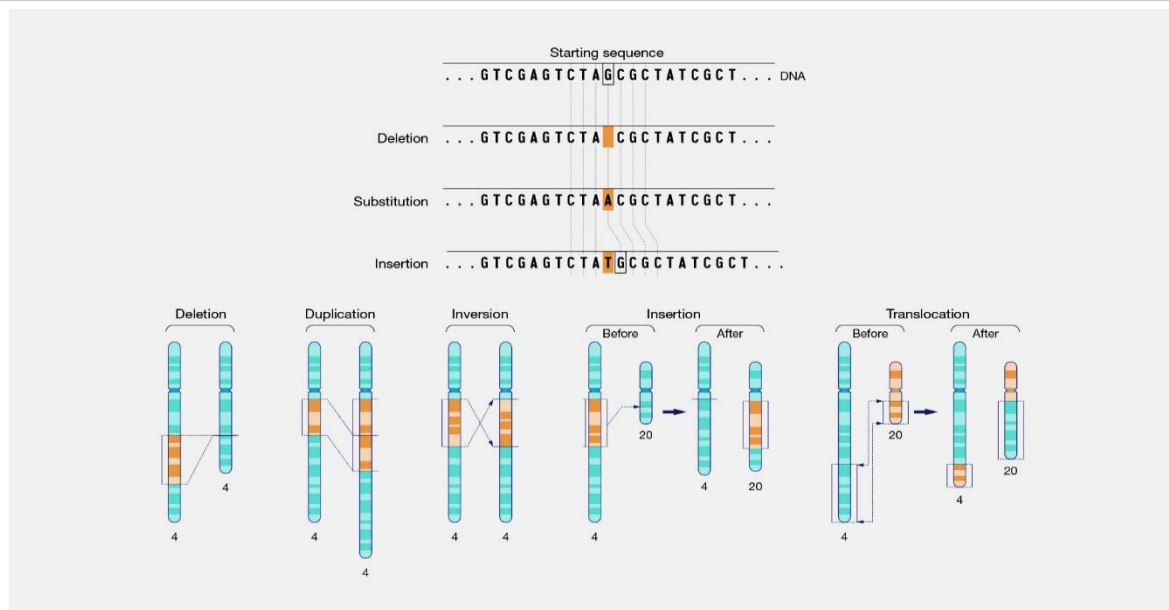


Figure 2: illustrates a point mutation in a gene (adapted from the National Human Genome Research Institute's Mutation webpage at <https://www.genome.gov/genetics-glossary/Mutation>).

AGTC are the four fundamental nucleotides that make up DNA, AGTC stand for Adenine, thymine, guanine and cytosine respectively [7]. Gene base changes can result in mutations, which can damage DNA and can be brought on by the inclusion, excision, or repetition of gene bases. Numerous variables, including as metabolic processes and environmental elements like radiation, can have an impact on DNA and cause thousands of distinct instances of DNA damage per cell each day[8]. Ability of the cell to copy the gene might be changed or lost when DNA is damaged. The process by which cells locate and repair DNA damage is known as DNA repair[9], and to respond to damage to the DNA structure, this process is ongoing. The disruption of cellular apoptosis may prevent the DNA damage from being repaired if natural repair mechanisms fail. According to the two-hit hypothesis, this irreparable damage can lead to the development of malignant tumors or cancer[10].

In this study, CRC is identified using an ensemble deep learning architecture. Three evaluation methods are employed: tenfold cross-validation testing, independent testing, and self-consistency testing. These techniques are applied to assess various traditional ML algorithms, such as SVM, GA-KASOM, DT (decision tree), and NV (Naive Bayes), RN (random forest) and DL algorithms such as CNN, BI-LSTM, GRU etc. Acc, Spe, Sen, ROC and MCC are the evaluation metrics after

applying the ML and DL algorithms

2 Literature review

Colorectal most cancers (CRC) stands because the fourth maximum not unusual reason of cancer-related deaths globally, with 1.4 million new cases and seven-hundred,000 fatalities suggested every year [11]. CRC traces back to ancient times, evidenced by the discovery of cancer in an Egyptian mummy from the Ptolemaic period, found in the Dakhleh Oasis, initially developing on the inner layer of the rectum or colon [12]. Depending on the type of polyp, there exists the potential for it to progress into cancer. Various types of adenomas (polyps) include Sessile serrated polyps (SSP), inflammatory polyps, adenomatous polyps (adenomas), and traditional serrated adenomas (TSA). Over time, cancerous polyps typically cause thickening of the colon or rectum wall layers. CRC originates and spreads within the mucosa, the deepest layer. The stages of CRC are determined by the depth of cancer cell infiltration into the wall and the extent of spread to nearby lymph nodes, blood vessels, or distant body areas. The majority of colorectal malignancies are adenocarcinomas, which originate in the cells that produce mucus to lubricate the colon and rectum [13].

Oncogenomic abnormalities drive the indefinite proliferation of cancer cells. While not every

mutation leads to cancer, not all cell proliferation results in malignancy. Cancer initiation arises from a disruption in the balance between cell creation and cell death due to alterations in gene function. Cancer driver genes are responsible for fueling cancer growth, often resulting in malignant mutations, unlike passenger gene mutations [14], [15]. Numerous studies have identified driver and passenger mutations based on clinical data. According to IntoGen, a renowned cancer driver gene database, there are 96 driver genes associated with colonic adenocarcinoma, a malignant tumor [16]. Bioinformatics is essential in the medical field, utilizing a range of computational and machine learning techniques for disease diagnosis and prevention. A study identified colorectal cancer subtypes after compiling a gene database linked to

CRC [14]. Multiple Machine Learning (ML) algorithms have been developed for colorectal adenocarcinoma detection and identification, with the accuracy of these processes reliant on Deep Learning and Artificial Intelligence (AI). Previous implementations include Kohonen Self-Organizing Map (KSOM) with genetic algorithms (GA) termed GA-KSOM, and the neuro-fuzzy algorithm, utilizing the Adaptive Neuro-Fuzzy Inference System (ANFIS) chosen as its structure/architecture. These yielded promising accuracies of 86.7% and 87.8%, respectively [17]. Additionally, a study utilizing a convolutional neural network (CNN) achieved sensitivity (Sen), specificity (Spe), and accuracy (Acc) rates of 67.5%, 89.0%, and 81.2%, respectively, with an AUC of 0.871 [18].

Table 1 shows the results of several studies and the accuracy of their proposed systems.

Table 1: Related Studies

Paper citation	Algorithm	Accuracy Achieved
21	GA-KSOM & ANFIS	87.80%
22	CNN	81.20%
23,24	ENDOSCOPIC NBI, CNN, SVM	90%
25	ENDOSCOPICAL DIP,	93.33%
26	RAMAN SPECTROSCOPY	98.50%
28	REGRESSION MODEL	73.9% TO 87.0%
28	RM-DM(DISCRIMINATION MODEL)	83.50%
30	FMA, FOLFOX	69.20%
31	Naïve Bayes	95.24%

Another study related to endoscopic imaging via deep learning by Okamoto et al. developed a diagnostic system for Narrow Band Imaging (NBI) using a CNN-based structure used as a feature extractor on the Bag-of-Features (BoF) and SVM for classification. In this way, the Accuracy level reached 90% [19][20]. Digital Image Processing (DIP) study incise by examining their histological pictures, a classification framework to distinguish between five kinds of colon tissues (two benign and three malignant) has been developed, 96.33% accuracy is achieved in this process [21]. Raman spectroscopy to detect CRC in this process with 98.5% accuracy is conducted [22]. Metastatic cancer spreads from its origin to other

body parts and has similar cancer cells. Another technique is to distinguish between metastatic and non-metastatic CRC, a support vector machine (SVM) classifier relying on the betweenness centrality (BC) technique was applied. This distinguishes non-metastatic CRC genes from metastatic. certain genes can potentially be biomarkers for the prognosis of metastatic CRC with significant accuracy [23].

In Addition, Study purposed to evaluate the decrease in the size of the tumor or the extent of cancer in any body part, for this Techniques used Regression Analysis. A logistic regression analysis and five-gene model were performed, and the resulting accuracies varied from 73.9 to 87.0

percent, with a 95 percent sensitivity and specificity. Moreover, the discriminative model showcased improvements, achieving an accuracy of 83.5%, sensitivity of 66.0%, specificity of 92.0%, a positive predictive value of 89.2%, and a negative predictive value of 74.0% [24].

Forest model with Decision Tree Aggregation is also used to detect Gut Microbes Associated with CRC, and decision tree DT is applied to diagnosing colorectal cancers. In this model, Significant accuracy has been seen [25]. Random Forests analysis is purposed to address responders to FOLFOX therapy which identified 80.0% sensitivity, 92.8% spec, and out-of-bag classification with 69.2% Acc [26] Naïve Bayes Classifier Models give 95.24% classification accuracy [27]. Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), and Long Short-Term Memory (LSTM) networks were employed to discern temporal patterns in EMR data, as

illustrated by the ROC Curve, resulting in an achieved AUC of 0.811 [28].

3 Material and methods

This section offers a comprehensive, step-by-step outline of data collection (datasets) and analysis employing diverse machine learning algorithms for cancer detection. Following this, subsequent sections delve deeply into the testing methodologies applied to these algorithms.

3.1 Methodology

In this work, colorectal cancer is identified using machine learning approaches. The study for the identification of cancer has used a number of machine learning methods. Figure 3 displays the proposed system's schematic diagram.

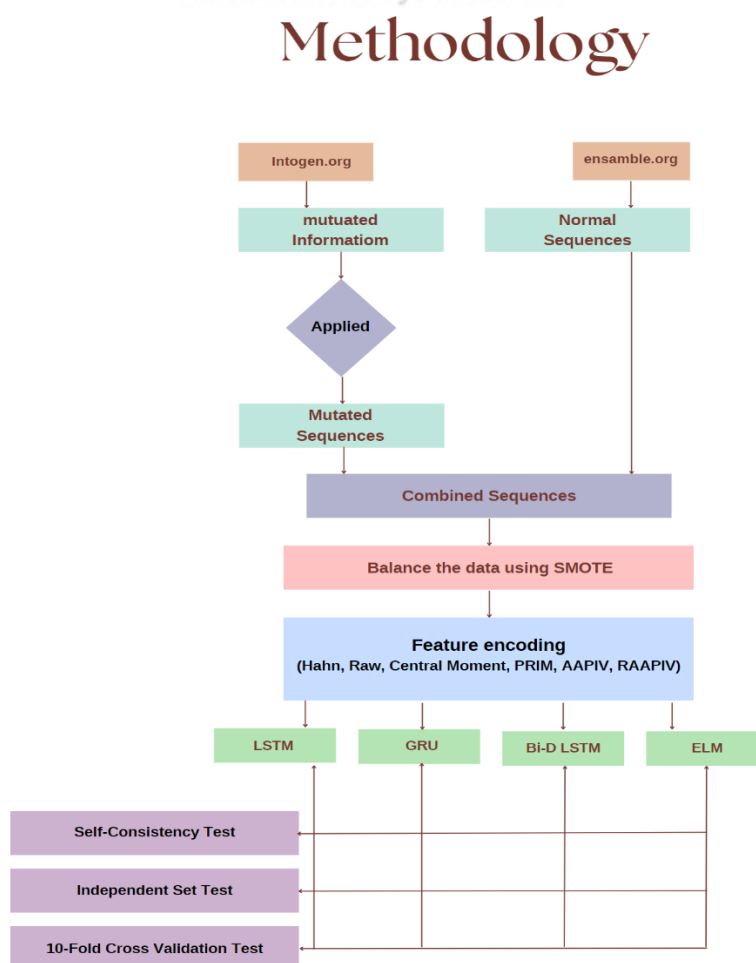


Figure 3: Methodology Diagram

Supervised and unsupervised machine learning methods are primarily utilized to classify deep

learning (DL) algorithms [29]. The DL framework was then categorized as part of the layer-based learning methodology. The input features are further processed by In each layer of the DL algorithm, information from the preceding layer is utilized as input. Features represent the pertinent information that each layer autonomously learns from the input data. [30]. By effectively resolving difficult challenges, DL frameworks represent a substantial advancement [30]–[32].

This study makes use of long term memory (LSTM), gated recurrent devices (GRUs), and bidirectional LSTM as deep getting to know algorithms to stumble on colorectal cancer. The efficacy of those fashions in diagnosing colorectal cancer is evaluated via three assessment metrics: the self-consistency test, the impartial set test, and the ten-fold cross-validation check.

3.1.1 Benchmark Datasets

The most important aspect of the research is the

datasets. These datasets are utilized for model training, model testing, and outcome validation. The dataset's structure ensures that it includes both normal and mutant sequences. The asia.ensembl.org [33] website is utilized in the suggested approach to collect the dataset of typical gene sequence. Information about mutated genes can be found at intogen.org [34]. web scrapping code (WSA) in python is used to extract data set from above sites because The database at intogen.org exclusively comprises data on mutant genes. To integrate this information with normal gene sequences retrieved from asia.ensembl.org, a Python application named Mutated Sequence Generator (MSG) was developed to generate mutated gene sequences. Passenger mutations are the name given to these mutated gene sequences[35]. In this instance, it is not referred to as an assist or a driver mutation. In this dataset we have 3,890 human samples and 5,693 mutations in total along with 72 distinct gene variants, as illustrated in table 2 below.

Table 2: genetic variations in genes

Symbol	Mutations	Samples	Symbol	Mutations	Samples
APC	1,055	785	SIRPA	55	11
TP53	749	712	ACVR2A	29	10
KRAS	514	515	FANCD2	26	10
PIK3CA	240	216	RBM10	27	10
SMAD4	144	139	ARID1B	20	10
FBXW7	158	130	MYH9	25	9
TCF7L2	91	98	TBX3	19	9
FAT4	196	83	NF1	28	9
AMER1	83	82	BMPRI1A	17	9
LRP1B	209	79	BCOR	17	9
BRAF	83	79	NIN	26	9
SOX9	49	69	USP6	29	9
ATM	84	66	EP300	22	8
NRAS	66	63	EGFR	18	8
ARID1A	47	52	FLCN	9	8
KMT2C	109	49	PTPN11	19	8
PCBP1	40	36	PARP4	19	7
EPHA3	40	35	EIF3E	8	6
BCL9L	43	34	MAP3K1	21	6
BCL9	36	30	BIRC6	51	6
SMAD2	34	29	NBEA	113	5

ERBB2	36	29	FAT3	167	5
SMAD3	30	29	CCDC6	6	4
PTEN	38	29	RNF6	14	4
RNF43	31	28	NCOR2	24	4
HSPG2	48	21	HRAS	7	4
CTNNB1	36	20	GRIN2A	66	3
BRCA2	42	19	WNK2	27	2
MTOR	36	16	PPP2R1A	16	2
GNAS	58	15	TGFBR2	21	2
PTPRC	35	15	MTCP1	2	2
SMARCA4	33	14	CARD11	65	2
MAP2K4	24	12	DICER1	20	2
DUSP16	18	11	BCORL1	28	2
MAP2K1	15	11	LATS1	17	2

3.1.2 Feature Extraction Framework

The input for the learning algorithms was quantitative. The proposed model incorporates a variety of fundamental quantitative encodings, dimensionality reduction strategies, and methods for extracting features. To obtain valuable features from the provided dataset for both training and

testing, a feature extraction framework (FEF) has been developed. The FEF will aid DL models in improving learning accuracy[36], [37]. In figure 4, the FEF operating sequence is displayed. To measure the many forms of information included in DNA sequences, it comprises of information-oriented statistical processes.

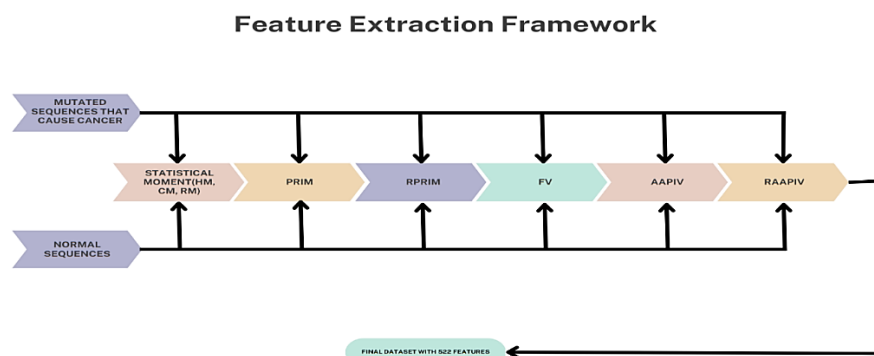


Figure 4: feature Extraction framework

3.1.2.1 Statistical moment

In this study, the general behavior and unpredictability of data can be assessed through statistical ordering. Specifically, statistical moments are employed to transform genetic data. The central movement (CN), raw movement (RM), and Hahn movement (HN) of genomic data

provide essential elements within the input vector for the predictor's use. Given that Hahn movement necessitates Genomic sequences are transformed into a two-dimensional notation, represented by a square matrix, to depict the data., K' of size $x * x$ as indicated in equation (1).

$$K = \begin{bmatrix} K_{11} & K_{12} \dots & K_{1n} \\ K_{21} & K_{22} \dots & K_{2n} \\ \vdots & \vdots & \vdots \\ K_{n1} & K_{n2} \dots & K_{nn} \end{bmatrix} \quad (1)$$

Hence K defines the sequence of gene [38]. Twofold is advantage of obtain matrix. The calculation of statistical moments for reducing dimensionality and the utilization of feature vectors with fixed sizes take precedence.

3.1.2.2 Hahn movement-HM

HMs are invertible and represent statistical homogeneity. Due to this inversion, HMs can be used to recreate the original data. The predictor makes use of the derived HM through the appropriate feature vector. The following Han polynomial is used to compute HM.

$$l_n^{r,s}(Q, R) = (R + W - 1)n(R - 1)n \times \sum_{z=0}^n (-1)^z (-1)^z \frac{(-1)_z (-n)_z (2R + r + s - n - 1)_z}{(R + s - 1)_z (Q - 1)_z} \frac{1}{z!} \quad (2)$$

$$H_{xy} = \sum_{i=0}^{H-1} \sum_{j=0}^{H-1} \delta_{xy} j_x^{a,b}(k, R) j_x^{a,b}(k, R), \quad m, n = 0, 1, 2, \dots, R - 1 \quad (3)$$

3.1.2.3 Raw movement-RM

Rm are differences in scale and space. The most typical kind of statistical moment is this one. It is used to figure out a probability distribution's mean, variance, and variance. Imputation is

performed using RM data. Imputation involves finding replacement values for missing data values in a dataset that are more likely to keep the information intact [39], [40].

Equation 4 describes how the RMs of 2D data of order $r + s$ are represented.

$$V_{rs} = \sum_{e=1}^n \sum_{f=1}^n e^r f^s \delta e f \quad (4)$$

To calculate the Central movement (CM), which serves as the centroid of the data, centroids (r, s) are necessary. The centroids may be used to calculate CMs as

$$U_{rs} = \sum_{e=1}^n \sum_{f=1}^n (e - \bar{x})^r (f - \bar{y})^s \delta e f \quad (5)$$

3.1.2.4 Relative Positional Incidence Matrix

The Relative Positional Incidence Matrix (RPIM) is instrumental in identifying the relative positions

of nucleotide bases. Mapping gene attributes requires careful consideration of both the base's location and its specific function. Equation is used to characterize it as a two-dimensional metric.

$$S_{PRIM} = \begin{bmatrix} S_{1 \rightarrow 1} & S_{1 \rightarrow 2} \dots & S_{1 \rightarrow q} \dots & S_{1 \rightarrow M} \\ S_{2 \rightarrow 1} & S_{2 \rightarrow 2} \dots & S_{2 \rightarrow q} \dots & S_{2 \rightarrow M} \\ \vdots & \vdots & \vdots & \vdots \\ S_{p \rightarrow 1} & S_{p \rightarrow 2} \dots & S_{p \rightarrow q} \dots & S_{p \rightarrow M} \dots \\ \vdots & \vdots & \vdots & \vdots \\ S_{M \rightarrow 1} & S_{M \rightarrow 2} \dots & S_{M \rightarrow q} \dots & S_{M \rightarrow M} \end{bmatrix} \quad (6)$$

In this case, $Rp \rightarrow q$ stands for the group of locations where the qth base is located relatively to the pth base's initial occurrence. After that, the matrix may be subjected to compute the HM, CM, and RMs.

3.1.2.5 Reverse position relative incidence matrix

The diversity in features obtained through permuting and rearranging gene sequences using specific statistical approaches always ensures the reliability. Equation (7) outlines the process for generating the reverse position relative abundance matrix (RPRIM) by reversing the original gene sequence. [25]. Additionally, HM, CM, and RMs are computed using RPRIM.

$$S_{PRIM} = \begin{bmatrix} S_{1 \rightarrow 1} & S_{1 \rightarrow 2} \dots & S_{1 \rightarrow q} \dots & S_{1 \rightarrow M} \\ S_{2 \rightarrow 1} & S_{2 \rightarrow 2} \dots & S_{2 \rightarrow q} \dots & S_{2 \rightarrow M} \\ \vdots & \vdots & \vdots & \vdots \\ S_{p \rightarrow 1} & S_{p \rightarrow 2} \dots & S_{p \rightarrow q} \dots & S_{p \rightarrow M} \dots \\ \vdots & \vdots & \vdots & \vdots \\ S_{M \rightarrow 1} & S_{M \rightarrow 2} \dots & S_{M \rightarrow q} \dots & S_{M \rightarrow M} \end{bmatrix} \quad (7)$$

3.1.2.6 Frequency vector

The frequency vector (FV) aids in the provision of compositional information and is used to Derive sequence correlations among nucleotide bases using the designated PRIM and RPRIM matrices.

The number of nucleotide occurrences throughout the gene sequence is determined by each FV component. DNA gene sequence structure is modelled using frequency matrices. Equation can be used to represent this (8).

$$\alpha = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n\} \quad (8)$$

In the DNA gene sequences, ε stand for the frequency of each nucleotide.

3.1.2.7 Accumulative absolute position incidence vector

Fuzzy patterns in DNA gene sequences may be found via feature extraction, which is particularly successful. Any nucleotide base in the gene

sequence can be found in the Accumulated Absolute Position Indices Vector (AAPIV), which offers cumulative information on positional occurrence. Equation (9) describes where the gene sequence is located.

$$L = \{\lambda_1, \lambda_2, \dots, \lambda_n\} \quad (9)$$

For any ith component,

$$\lambda_i = \sum_{k=1}^n \beta_k \quad (10)$$

3.1.2.8 Reverse accumulative absolute position incidence vector (RAAPIV)

Finding the many and intricate patterns in DNA gene sequences is made easier by reverse sequencing. Similar to how AAPIV operates, RAAPIV makes educated guesses about the crucial facts in the opposite order. The following is the RAAPIV equation (11):

$$\lambda = \{n_1, n_2, \dots, n_m\} \quad (11)$$

In λ , each element represents the total count of times a nucleotide appears in the reverse sequence. Each of the statistical operations mentioned above that are utilized for feature extraction has a distinct biological relevance [41], [42]. These operations glean details about the location and composition of the DNA gene sequence. Characteristics such as the occurrence frequency of individual elements within a DNA gene sequence and their respective positional frequencies, the structure of specific genes, and the absolute positioning of each element within a gene sequence can all be extracted through feature extraction techniques.

3.1.3 Prediction Algorithms:

Deep Learning (DL) algorithms encompass both supervised and unsupervised machine learning techniques [43]. The DL architecture is recognized as a learning methodology comprising multiple layers. Within the DL algorithm, each layer receives input from the preceding layer and conducts further processing of input features. These features encapsulate relevant information for each layer, enabling the DL algorithm to glean insights from the input data [30]. Deep Learning

Frameworks represent a substantial advancement in effectively addressing challenging problems [31], [32], [44].

For this study, Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs), Bidirectional LSTM, and an Ensemble Learning Model (ELM) are utilized to identify colorectal cancer (CRC). To assess CRC detection, the models are subjected to testing using three metrics: SCT, IST, and TCVT.

3.1.3.1 LSTM network

LSTM stands as a state-of-the-art variant of Recurrent Neural Networks (RNNs), effectively addressing the vanishing gradient problem inherent in traditional RNNs. Employing gates to manage the reading, writing, and storage of data, LSTM represents a gating mechanism. These gates—namely input gates, forget gates, and output gates—operate sequentially. The forget gate decides if the information should be stored in or forgotten from the current memory state, enabling the movement of hidden memory states to the immediate next. It is also responsible for the suppression of facts in next processing steps. Considering the range of activation features, input gate is for changing the country of the mobile and output gate for assaying the output produced from modern cell state [45]. Figure 5 provides a description of the LSTM layer model employed in this investigation.

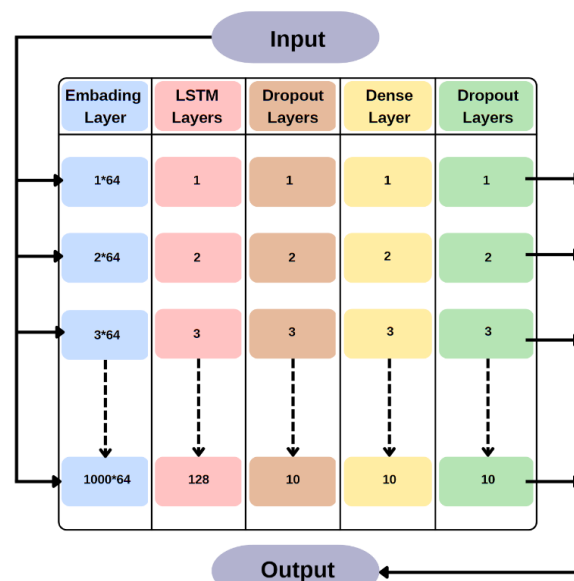


Figure 5: Layered LSTM model employed for colorectal cancer identification.

The input undergoes transformation into a fixed-length vector of specified size via the embedding layer within the LSTM. With word vectors of 64 characters and a vocabulary capacity of 1000, the subsequent layer consists of an LSTM layer featuring an output of 128 neurons. Following this, two dropout layers are incorporated, each inhibiting 10% of neurons to prevent overfitting. A dense layer comprising ten neurons follows suit. Stochastic Gradient Descent (SGD) serves as the

optimization method within the LSTM framework. Additionally, the Sparse Categorical Cross Entropy (SCCE) function is utilized as a loss minimizer, accompanied by sigmoid activation function.

The gate manages the regulation of information flow between cells in the LSTM. Each gate applies activation functions differently through [47]. The diagram in Figure 6 depicts how the gates function within the LSTM model [48].

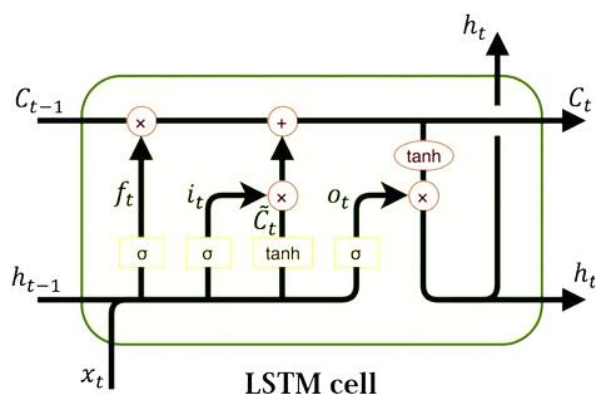


Figure 6: A simple LSTM cell Structure.

where x_t is the input and h_t is the output at specific time t . f_t represents forget gate, and i_t represents input gate and output gate respectively. Every cell of LSTM has three inputs and has two outputs as. Equation 12, 13, 14, 15, 16, 17 explain LSTM Where h_t is the output at a certain moment and

x_t is the input Forget gate is represented by f_t whereas input gate and output gate are represented by i_t and o_t respectively. Each LSTM cell contains three inputs x_t, h_{t-1}, C_{t-1} and two outputs h_t and C_t . LSTM is explained in equations 12, 13, 14, 15, 16, and 17.

$$i_t = \sigma(x_t V^i + h_{t-1} U^i) \quad (12)$$

$$f_t = \sigma(x_t V^f + h_{t-1} U^f) \quad (13)$$

$$o_t = \sigma(x_t V^o + h_{t-1} U^o) \quad (14)$$

$$\tilde{C}_t = \tanh(x_t V^g + h_t U^g) \quad (15)$$

$$\tilde{C}_t = \sigma(f_t * C_{t-1} + i_t * \tilde{C}_t) \quad (16)$$

$$h_t = \tanh(\tilde{C}_t) * o_t \quad (17)$$

where h_{t-1} is the previous data cell output, C_{t-1} is the previous cell memory, C_t is the current cell memory? While V and U are the weights for the forget, input, and output gate.

where C_{t-1} is the previous cell memory, C_t is the current cell memory, and h_{t-1} is the prior data cell output? The input and output gates' weights are V and U , respectively.

3.1.3.2 Gated recurrent unit (GRU)

In the proposed research, another DL algorithm utilized is the GRU [1]. It functions similarly to LSTM but with fewer parameters and gates. Despite having less training data, GRU outperforms LSTM. Additionally, opting for GRU over LSTM is advantageous due to its reduced complexity and superior performance. GRU solely employs the reset gate and update gate within its cell [45]. Figure 7 explains the layered structure of GRU in the diagnosis of CRC.

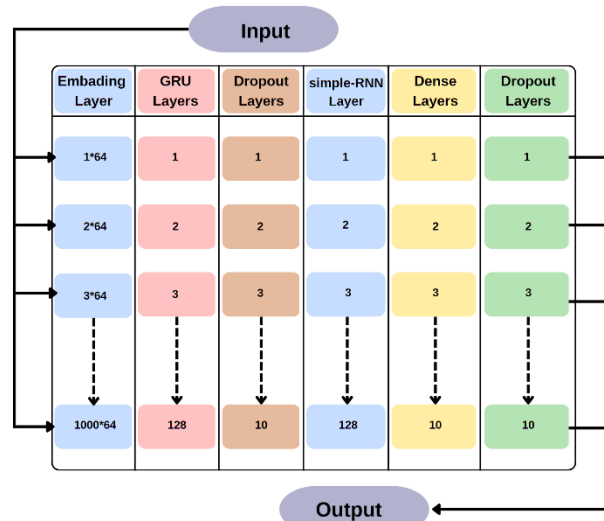


Figure 7: GRU layered structure model for identification of CRC.

The GRU architecture recommended is with an embedding layer, that convert the input to a vector of dimension size of your choice. This layer has a 64-character word vector length and is able to support a vocabulary of 1000 words. And then a GRU layer with 256 neurons in output. We also added a RNN layer of a single layer with 128

neurons and 2 dropouts to tackle overfitting by deactivating 10% of neurons. Dense (10) is applied as the activation function and the optimizer within the GRU model is SGD. Activation is performed using the sigmoid function, and the SCCE function is utilized to minimize loss.

In the GRU architecture, the reset gate determines the extent to which historical data is disregarded, while the update gate decides the

degree to which historical data is utilized. The structural configuration of the GRU cell is illustrated in Figure 8. [49]

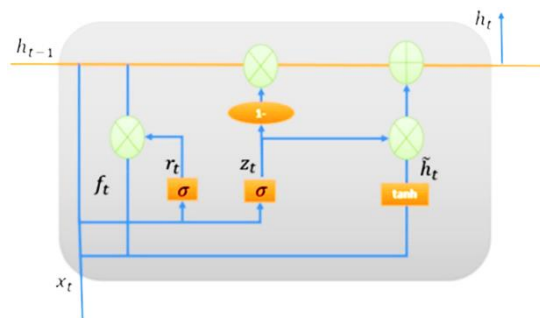


Figure 8: The basic structure of a GRU cell, which stands for Gated Recurrent Unit.

Equations 18,19, 20, and 21 describe how the GRU operates.

$$r_t = \sigma(x_t A^r + h_{t-1} - B^r) \quad (18)$$

$$z_t = \sigma(x_t A^z + h_{t-1} - B^z) \quad (19)$$

$$h'_t = \tanh(r_t * h_{t-1} A + x_t B) \quad (20)$$

$$h_t = (1 - z_t) * h'_t + z_t * h_{t-1} \quad (21)$$

In equations (18) and (19), r_t stands for the reset gate and z_t for the update gate, respectively. In comparison to LSTM, GRU requires less computing time[50]

3.1.3.3 Bi-Directional LSTM

The method chosen in the study is DL bidirectional LSTM [51]. The model consists of two LSTM units connected to an output layer, one operating in the forward direction and the other operating in the reverse direction. Fixed length vector of the dimension. The word vector is 64 characters long and the word capacity is 1000. Three levels of waste were also compiled. To solve the overfitting problem, the first intervention layer affects 20% of the neurons, the second layer affects 20%

, and the third layer affects 10%. Two layers are then added; The first has 64 neurons, the second has 10 neurons. The bidirectional LSTM model uses adaptive moment estimation (ADAM) as the optimizer and starts using the sigmoid function. Binary cross entropy function is used to Reduce loss.

This mechanism learns independently without requiring prior information, as it simultaneously processes both forward and backward information [52]. In Figures 9 and Figure 10, the suggested Bi-directional LSTM is described.

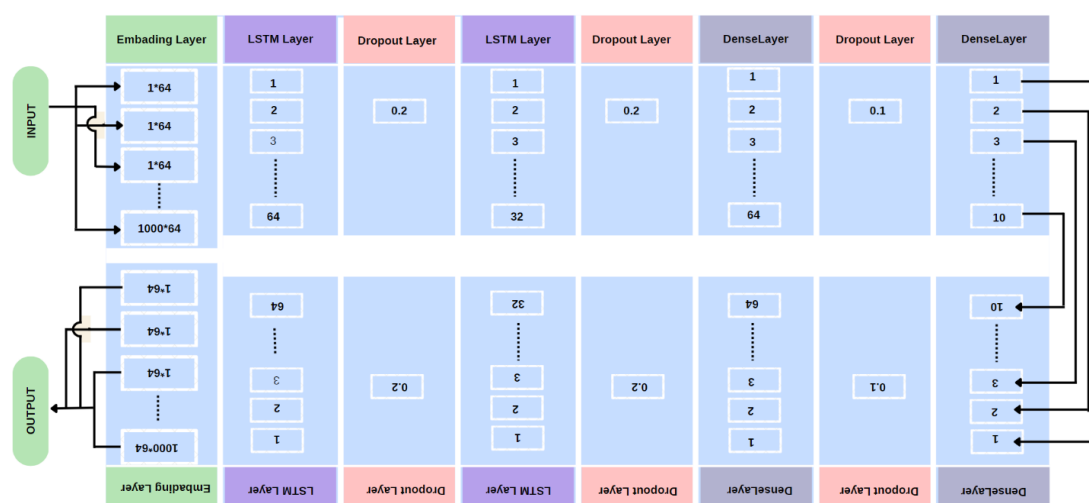


Figure 9: Bi-directional LSTM model showcases bidirectional processing, utilizing LSTM's long and short-term memory capabilities.

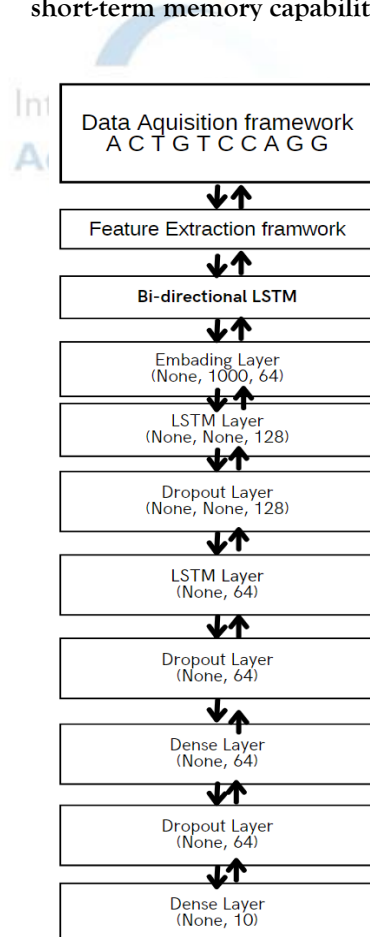


Figure 10: Bi-directional LSTM structure for colorectal cancer detection, utilizing LSTM for memory processing.

3.1.3.4 Ensemble Learning Model (ELM)

The ELM employs the Divide and Conquer (DAC) approach to enhance the accuracy of individual base learners before consolidating the entire

model. To achieve optimal performance, numerous basic learners are integrated [53]. Each base learner produces specific results, which are aggregated and used to learn various features

through data chunks obtained via the bootstrap technique. This allows the model to extract hidden patterns from datasets. Ensemble learning, a versatile technique, outperforms simple machine learning algorithms in accuracy because it trains a model using all possible combinations of data and a bootstrap approach, which allows for element and row replacement procedures. Additionally, this approach solves overfitting problems. Bagging [54], boosting [55], and stacking [56] are three prominent types of ensemble learning models, each aimed at achieving high accuracy. Among these models, we will adopt the Stacking (Stacking Model) approach, which involves the combination of heterogeneous machine learning models (basic learners) using a different data mining technique [57]. Initially, the core learners are trained, followed by the instruction of a combinatory also referred to as a meta-classifier, to produce a final prediction based on the core learners' predictions [58]. Stacked ensembles outperform any single

elementary learner such as a single RF or GBM and are shown to be an optimally efficient learning system. [59], [60]. Employing an ensemble learning approach, this research improves the performance of various deep learning methods such as LSTM, GRU, and bidirectional LSTM. The completed process is divided into three parts: training, validation (symbol V) and testing (symbol T). All deep learning models, including LSTM, GRU, and bidirectional LSTM, are trained using training data. Additionally, grid search optimization is used to determine the search and determine the best value for the ensemble learning model. Trained learning models are developed for each different deep learning model, denoted as trained model1, trained model2, and trained model3 for LSTM, GRU, and two-way LSTM, respectively, as shown in Figure 11.

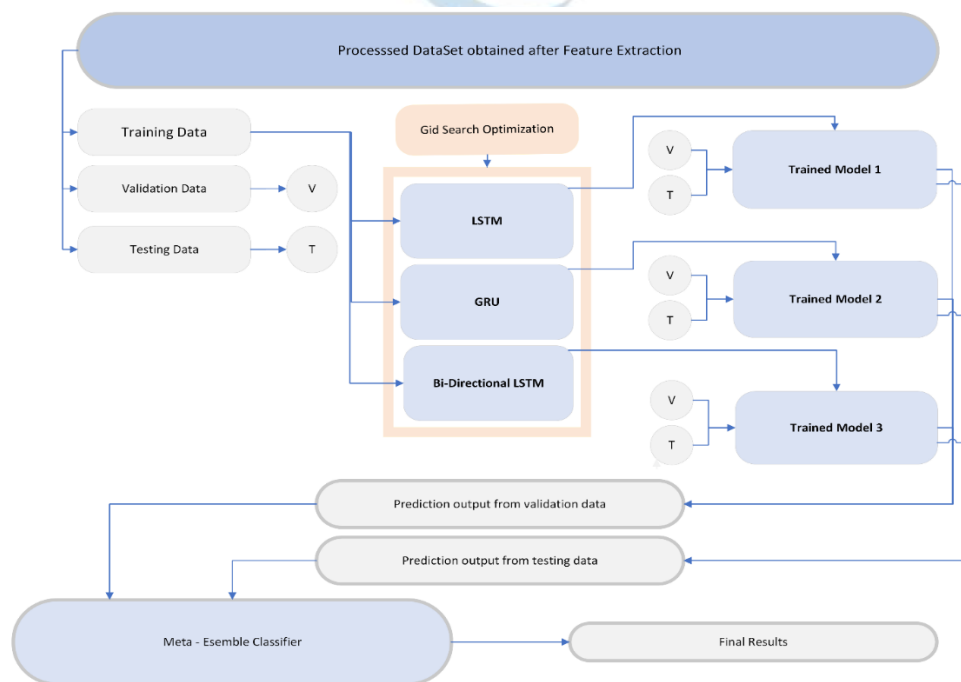


Figure 11: THE PROPOSED ENSEMBLE LEARNING MODEL OF GRU, BI-DIRECTIONAL LSTM AND LSTM APPROACH

All trained models undergo validation and testing using both validation and testing datasets. Finally, as demonstrated in the results section, an ensemble learning model produces ultimate enhanced outcomes.

$$yq, k = \sum_{o=1}^W Mof_{o,k} \quad (22)$$

In the equation, Each individual deep learning model is assigned weights to contribute to the ensemble learning prediction. Here, $M_o (o = 1, 2, 3, 4, \dots, W)$ Each deep learning model is allocated a weight $f_{o,k}$. It indicates the prediction made by each individual deep learning model and matches the observation. The models are trained for 20 epochs. Evaluation of DL approaches includes metrics such as AUC, precision, F1 score, recall, Cohen's kappa, specificity, sensitivity, Matthews correlation coefficient, loss, and precision, calculated in 20 iterations of forward and backward processes. [61], [62] Sen, Spe, Acc, and MCC are each described by equations (23), (24), (25) and (26), respectively.

$$\text{Sensitivity: } Sen = \frac{YO}{YO + GB} \quad (23)$$

$$\text{Specificity: } Spe = \frac{YB}{YB + GO} \quad (24)$$

$$\text{Accuracy: } Acc = \frac{YO + YB}{YO + GO + YB + GB} \quad (25)$$

$$MCC = \frac{(YO \times YB) - (GO \times GB)}{\sqrt{(YO + GO)(YO + GB)(YB + GO)(YB + GB)}} \quad (26)$$

Within these equations:

YB = each and every true negative values

YO = All of the dataset's true positive values

GB = erroneous negative values

GO = Falsely optimistic values

In the proposed research, Sensitivity pertains to the ability of a test to accurately identify cases of CRC, while Specificity refers to its ability to accurately identify individuals without CRC in the dataset [63]. All individuals with a specific condition are characterized by $YO + GB$ in the equations. $YB + GO$, however, represents all the subjects without the specified requirements. The total number of patients who had positive results is $YO + GO$, while the number of subjects who received negative findings is $YB + GB$ [64]. The procedure may be summarized using a confusion matrix [65]–[69]. In the confusion matrix, the entries YB stand for True Negative, YO for true positive values, GB for erroneous negative and GO for Falsely optimistic stand for True Negative, False Negative, and False Positive values, respectively.

4 Results and discussion

Before being input into deep learning (DL) techniques, the CRC dataset undergoes preprocessing to extract fundamental components and ensure balanced data. Subsequently, the processed data undergoes verification using the independent set test, self-consistency test, and 10-fold cross-validation test to assess the performance of DL algorithms. The results obtained from these validation methods are detailed in this section.

4.1 Self-Consistency test (SCT)

The DL algorithm is evaluated using the SCT technique, where all data is utilized for both training and assessment. In this method, the entire dataset is employed for both training and testing. While Bidirectional LSTM showcases minimal loss, LSTM, GRU, and Bidirectional LSTM achieve notably high accuracy in the SCT, as demonstrated in Table 3. Figure 12 depicts the decision boundary (DB) of SCT and Figure 13 shows the ROC Curve.

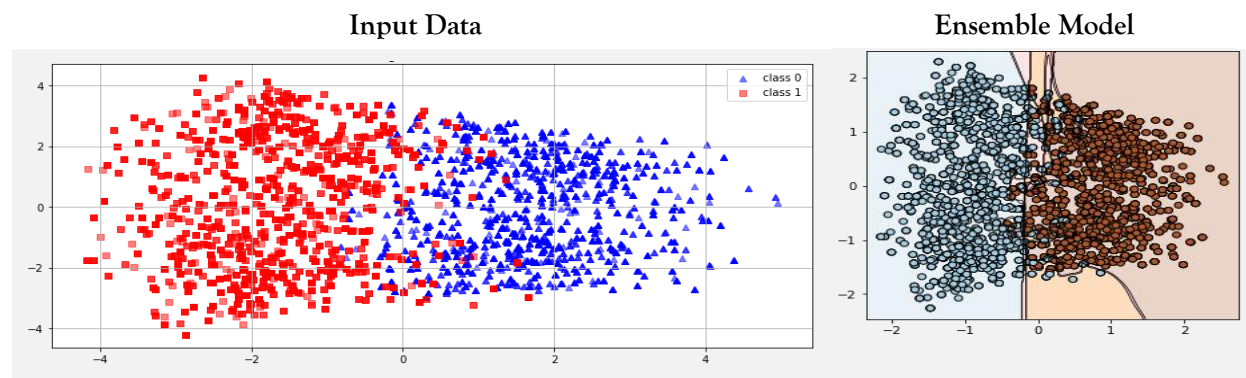


Figure 12: DB of SCT in EDLM

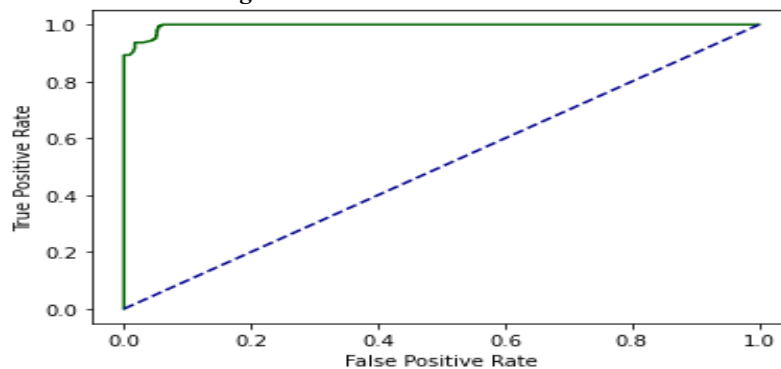


Figure 13: SCT's Roc curve

4.2 Independent Set test (IST)

The second testing method utilized for the proposed EDLM strategy is IST. The model's accuracy is determined by extracting values from

the misperception matrix. IST is the primary method proposed for assessing the model's performance. 20% of the dataset's values are used for testing, while the remaining 80% are used to train the algorithm.

Figure 14 depicts the decision boundary (DB) of IST and Figure 15 shows the ROC Curve.

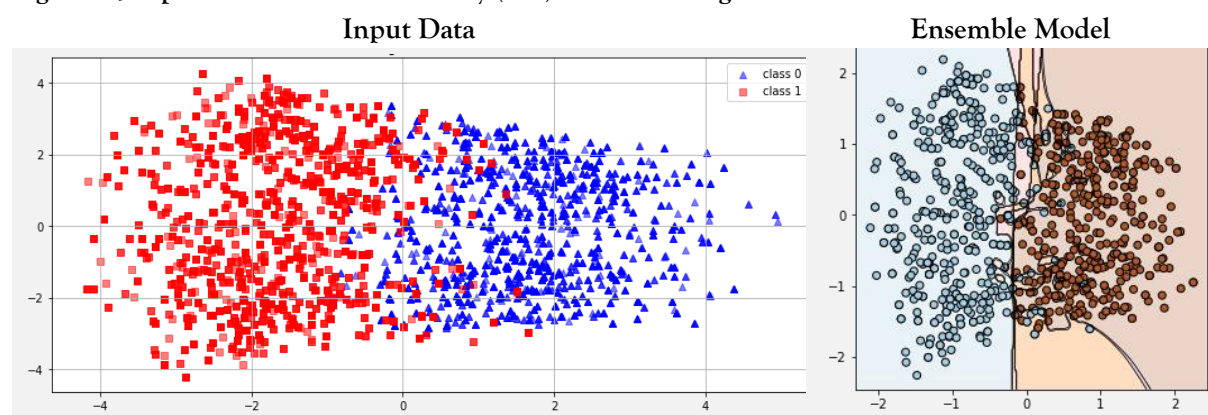


Figure 14: DB of IST in EDLM

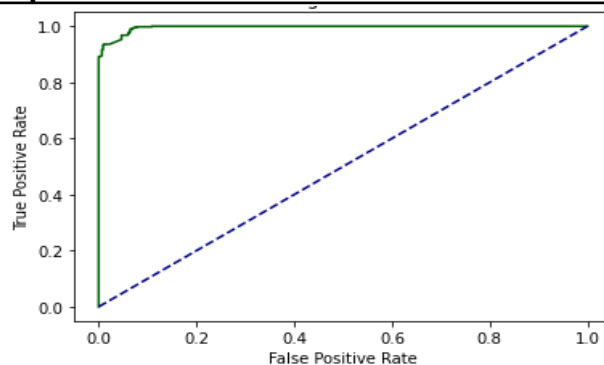


Figure 15: displays the IST's Roc curve.

4.3 10-Fold Cross validation test(10-FCVT)

The 10-fold cross-validation (FCV) technique involves dividing the data into ten equal subsets. Following model training, each subset of the training data is treated as a separate validation set.

Hyper-parameter tuning and architectural decisions are subsequently based on the overall performance across the ten folds. Figure 16 illustrates the decision boundary (DB) of the 10-FCVT.

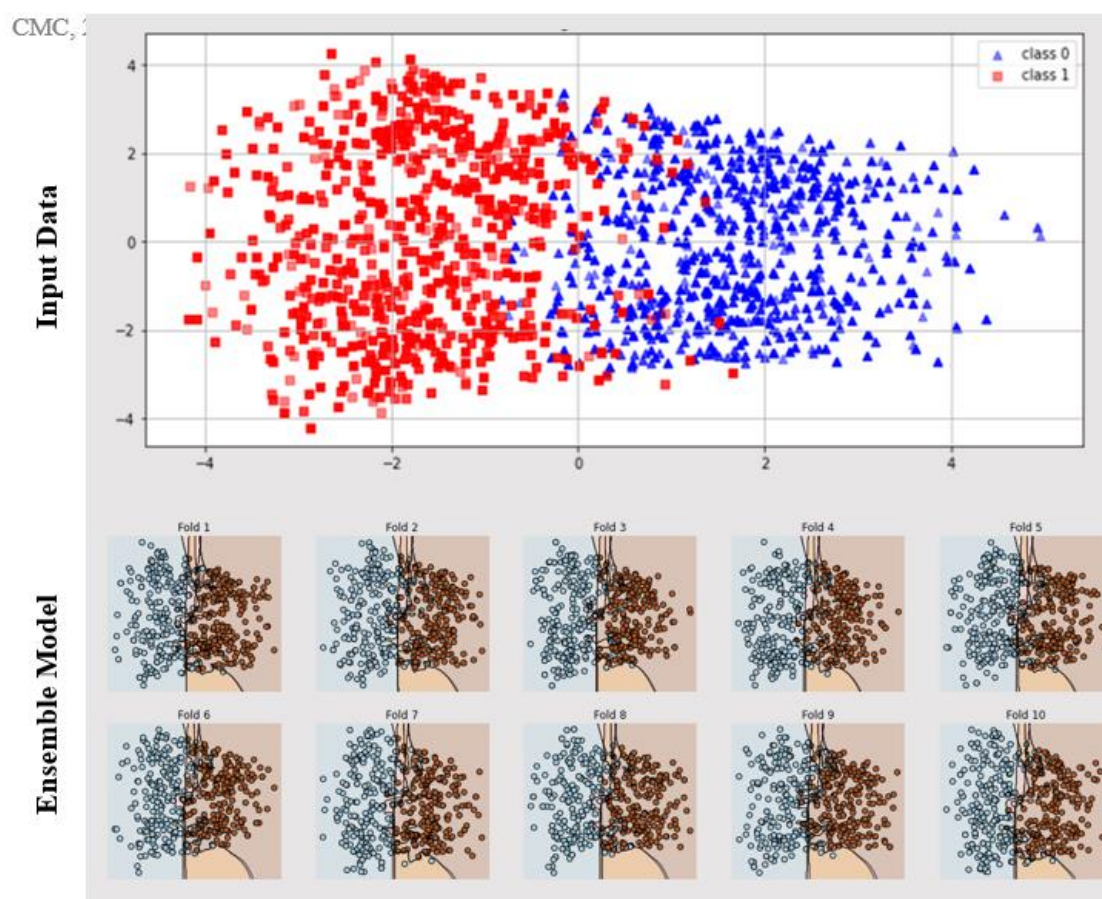


Figure 16: DB of 10-FCVT

Figure 17 below shows the ROC curve of the EDLM in 10-FCVT for all DL algorithms, including LSTM, GRU, and Bi-directional LSTM.

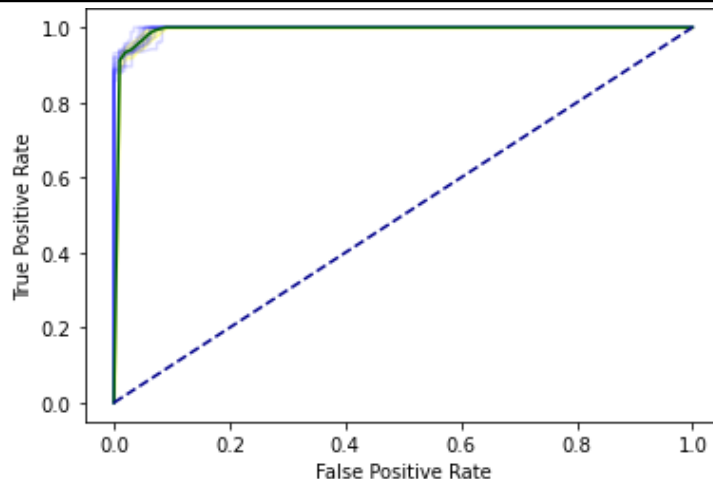


Figure 17: 10-FCVT'S ROC CURVE

4.4 Result Comparison

Below is Table 3, which consolidates the outcomes and juxtaposes the algorithms (LSTM, GRU, BLSTM, and EDLM) against SCT, IST, and 10-FCVT.

Table 3: Results of SCT, IST, and 10-FCVT concerning LSTM, GRU, Bi-LSTM, and EDLM are presented below.

	Self-Consistency Test				Independent Set Test				10-Fold Cross-Validation Test			
	Acc (%)	Sen (%)	Spe (%)	MCC (%)	Acc (%)	Sen (%)	Spe (%)	MCC (%)	Acc (%)	Sen (%)	Spe (%)	MCC (%)
LSTM	97	97	98	0.94	96	95	97	0.92	94	95	94%	0.88
GRU	96	95	97	0.92	95	95	96	0.91	94	94	93%	0.88
Bi-LSTM	99	99	100	0.98	98	97	99	0.97	94	95	94%	0.88
Ensemble	95	94	96	0.90	95	95	95	0.91	96	96	95%	0.91

Table-3 illustrates the efficacy of four machine learning models (LSTM, GRU, Bi-LSTM, and Ensemble) throughout 3 evaluation methodologies: the Self-Consistency test (SCT), the impartial Set test (IST), and the ten-Fold cross-Validation test (10-FCVT). The assessment criteria encompass Accuracy, Sensitivity, Specificity, and MCC.

Accuracy quantifies the ratio of correct predictions made by the model, while Sensitivity gauges its capability to accurately recognize positive cases. Specificity examines the model's aptitude to accurately identify negative cases, and MCC provides a holistic assessment of binary classification quality, taking into account true and false positives and negatives.

According to the results, the top-performing model is the Bi-LSTM, demonstrating the highest

accuracy, sensitivity, specificity, and MCC across most tests. For instance, in the Self-Consistency Test, For the Bi-LSTM, an accuracy of 99%, sensitivity of 99%, specificity of 100%, and MCC of 0.98 are attained. indicating that it performed exceptionally well. The Ensemble model performed well overall, but was outperformed by the Bi-LSTM in most of the tests. The LSTM and GRU models also performed well, but with slightly lower results compared to the other two models.

In conclusion, the Bi-LSTM model appears to be the best performing model based on the results in the table 3.

5 Conclusion

The research presents an Ensemble Deep Learning Model (EDLM) designed for the early

identification of colorectal adenocarcinoma, a particularly aggressive form of cancer impacting the colon and rectum. Employing deep learning methodologies like GRU, LSTM, and Bi-LSTM, the model evaluates normal and mutant gene sequences sourced from diverse outlets. Various feature extraction techniques are employed to transform the data into a quantitative format conducive to training and assessment. The findings of the study demonstrate that the proposed EDLM model achieves exceptional accuracy, boasting an overall accuracy rate of 96%. The model underwent assessment through three awesome methods: the 10-Fold Cross-Validation check (10-FCVT), the Self-Consistency test (SCT), and the independent Set test (IST), revealing encouraging effects, significantly excelling inside the IST. furthermore, comparative exams had been conducted against alternative algorithms, such as LSTM, GRU, and Bi-LSTM. overall performance metrics along with accuracy, sensitivity, specificity, and Matthew's correlation coefficient were computed for thorough evaluation. The findings of the study imply the efficacy of the proposed EDLM model for early most cancers detection. However, it should be noted that the model has only been validated for colorectal adenocarcinoma and more research is required to investigate its potential use in detecting other life-threatening diseases and to compare the effectiveness of various deep learning techniques for further improvement in accuracy and efficiency.

Funding Statement:

The authors received no specific funding for this study.

Data Availability:

The genomic data that support the findings of this study are openly available at <https://www.intogen.org/search?cancer=COAD> READ and http://asia.ensembl.org/Homo_sapiens/Gene/Sequence?g=ENSG00000134982;r=5:112707518-112846239.

Conflict of interest statement:

The authors declare that they have no conflicts of

interest.

REFERENCES

- [1] H. Brenner, M. Kloor, and C. P. Pox, "Colorectal cancer," *The Lancet*, vol. 383, no. 9927, pp. 1490–1502, Apr. 2014, doi: 10.1016/S0140-6736(13)61649-9.
- [2] S. P. Nguyen, S. Bent, Y. H. Chen, and J. P. Terdiman, "Gender as a risk factor for advanced neoplasia and colorectal cancer: a systematic review and meta-analysis," *Clin Gastroenterol Hepatol*, vol. 7, no. 6, 2009, doi: 10.1016/J.CGH.2009.01.008.
- [3] C. Stock, D. Pulte, U. Haug, and H. Brenner, "Subsite-specific colorectal cancer risk in the colorectal endoscopy era," *Gastrointest Endosc*, vol. 75, no. 3, pp. 621-630.e1, Mar. 2012, doi: 10.1016/J.GIE.2011.10.025.
- [4] M. M. Center, A. Jemal, R. A. Smith, and E. Ward, "Worldwide Variations in Colorectal Cancer," *CA Cancer J Clin*, vol. 59, no. 6, pp. 366–378, Nov. 2009, doi: 10.3322/caac.20038.
- [5] G. J. Poston, D. Tait, S. O'Connell, A. Bennett, and S. Berendse, "Diagnosis and management of colorectal cancer: summary of NICE guidance," *BMJ*, vol. 343, no. 7831, pp. 1010–1012, Nov. 2011, doi: 10.1136/BMJ.D6751.
- [6] D. M. Fitzgerald and S. M. Rosenberg, "What is mutation? A chapter in the series: How microbes 'jeopardize' the modern synthesis," *PLoS Genet*, vol. 15, no. 4, p. e1007995, 2019, doi: 10.1371/JOURNAL.PGEN.1007995.
- [7] S. Tolosa, J. A. Sansón, and A. Hidalgo, "Theoretical study of adenine to guanine transition assisted by water and formic acid using steered molecular dynamic simulations," *Front Chem*, vol. 7, no. JUN, p. 414, Jun. 2019, doi: 10.3389/FCHEM.2019.00414/BIBTEX.

- [8] S. P. Jackson and J. Bartek, "The DNA-damage response in human biology and disease," *Nature* 2009 461:7267, vol. 461, no. 7267, pp. 1071–1078, Oct. 2009, doi: 10.1038/nature08467.
- [9] A. E. Pegg, "Multifaceted roles of alkyltransferase and related proteins in DNA repair, DNA damage, resistance to chemotherapy, and research tools," *Chem Res Toxicol*, vol. 24, no. 5, pp. 618–639, May 2011, doi: 10.1021/TX200031Q/ASSET/IMAGES/MEDIUM/TX-2011-00031Q_0004.GIF.
- [10] X. Zhu, H. gon Lee, G. Perry, and M. A. Smith, "Alzheimer disease, the two-hit hypothesis: An update," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1772, no. 4, pp. 494–502, Apr. 2007, doi: 10.1016/J.BBADIS.2006.10.014.
- [11] X. Zhu, A. K. Raina, G. Perry, and M. A. Smith, "Alzheimer's disease: the two-hit hypothesis," *Lancet Neurol*, vol. 3, no. 4, pp. 219–226, Apr. 2004, doi: 10.1016/S1474-4422(04)00707-0.
- [12] J. Ferlay et al., "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *Int J Cancer*, vol. 136, no. 5, pp. E359–E386, Mar. 2015, doi: 10.1002/IJC.29210.
- [13] A. Rehemtulla, "Dinosaurs and Ancient Civilizations: Reflections on the Treatment of Cancer," *Neoplasia*, vol. 12, no. 12, pp. 957–968, Dec. 2010, doi: 10.1593/NEO.101588.
- [14] "What Is Colorectal Cancer? | How Does Colorectal Cancer Start?" <https://www.cancer.org/cancer/colon-rectal-cancer/about/what-is-colorectal-cancer.html> (accessed Jan. 14, 2023).
- [15] "Biopsy | Cancer.Net." <https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/tests-and-procedures/biopsy> (accessed Jan. 14, 2023).
- [16] J. R. Pon and M. A. Marra, "Driver and Passenger Mutations in Cancer," <https://doi.org/10.1146/annurev-pathol-012414-040312>, vol. 10, pp. 25–50, Jan. 2015, doi: 10.1146/ANNUREV-PATHOL-012414-040312.
- [17] L. A. Gan Lim, R. N. G. Maguib, E. P. Dadios, and J. M. C. Avila, "Implementation of GA-KSOM and ANFIS in the classification of colonic histopathological images," *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 2012, doi: 10.1109/TENCON.2012.6412240.
- [18] N. Ito, H. Kawahira, H. Nakashima, M. Uesato, H. Miyauchi, and H. Matsubara, "Endoscopic Diagnostic Support System for cT1b Colorectal Cancer Using Deep Learning," *Oncology*, vol. 96, no. 1, pp. 44–50, Dec. 2019, doi: 10.1159/000491636.
- [19] Z. Cao et al., "A Deep Learning Approach for Detecting Colorectal Cancer via Raman Spectra," *BME Front*, vol. 2022, pp. 1–10, May 2022, doi: 10.34133/2022/9872028.
- [20] M. Masud, N. Sikder, A. al Nahid, A. K. Bairagi, and M. A. Alzain, "A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework," *Sensors (Basel)*, vol. 21, no. 3, pp. 1–21, Feb. 2021, doi: 10.3390/S21030748.

- [21] J. Zhi, J. Sun, Z. Wang, and W. Ding, "Support vector machine classifier for prediction of the metastasis of colorectal cancer," *Int J Mol Med*, vol. 41, no. 3, pp. 1419–1426, Mar. 2018, doi: 10.3892/IJMM.2018.3359/HTML.
- [22] I. Pacal, D. Karaboga, A. Basturk, B. Akay, and U. Nalbantoglu, "A comprehensive review of deep learning in colon cancer," *Comput Biol Med*, vol. 126, p. 104003, Nov. 2020, doi: 10.1016/J.COMPBIOMED.2020.104003.
- [23] T. Okamoto et al., "Feature extraction of colorectal endoscopic images for computer-aided diagnosis with CNN," 2019 2nd International Symposium on Devices, Circuits and Systems, ISDCS 2019 - Proceedings, May 2019, doi: 10.1109/ISDCS.2019.8719104.
- [24] Y. T. Chang et al., "Gene expression profile of peripheral blood in colorectal cancer," *World Journal of Gastroenterology : WJG*, vol. 20, no. 39, p. 14463, Oct. 2014, doi: 10.3748/WJG.V20.I39.14463.
- [25] D. Ai, H. Pan, R. Han, X. Li, G. Liu, and L. C. Xia, "Using Decision Tree Aggregation with Random Forest Model to Identify Gut Microbes Associated with Colorectal Cancer," *Genes* 2019, Vol. 10, Page 112, vol. 10, no. 2, p. 112, Feb. 2019, doi: 10.3390/GENES10020112.
- [26] S. Tsuji et al., "Potential responders to FOLFOX therapy for colorectal cancer by Random Forests analysis," *British Journal of Cancer* 2012 106:1, vol. 106, no. 1, pp. 126–132, Nov. 2011, doi: 10.1038/bjc.2011.505.
- [27] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," *IOP Conf Ser Mater Sci Eng*, vol. 546, no. 5, p. 052068, Jun. 2019, doi: 10.1088/1757-899X/546/5/052068.
- [28] R. Amirkhan, M. Hoogendoorn, M. E. Numans, and L. Moons, "Using recurrent neural networks to predict colorectal cancer among patients," 2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings, vol. 2018-January, pp. 1–8, Feb. 2018, doi: 10.1109/SSCI.2017.8280826.
- [29] Y. Bengio, Y. Lecun Presenter, and M. Vazquez, "Scaling Learning Algorithms Towards AI," 2017.
- [30] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* 2015 521:7553, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [31] R. Vinayakumar, M. Alazab, S. Srinivasan, Q. V. Pham, S. K. Padannayil, and K. Simran, "A Visualized Botnet Detection System Based Deep Learning for the Internet of Things Networks of Smart Cities," *IEEE Trans Ind Appl*, vol. 56, no. 4, pp. 4436–4456, Jul. 2020, doi: 10.1109/TIA.2020.2971952.
- [32] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," *IEEE Access*, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.
- [33] "Gene: TP53 (ENSG00000141510) - Summary - Homo_sapiens - Ensembl genome browser 108." https://asia.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000141510;r=17:7661779-7687538 (accessed Jan. 14, 2023).

- [34] "IntOGen - Cancer driver mutations in Colorectal adenocarcinoma." <https://intogen.org/search?cancer=COR> EAD (accessed Jan. 14, 2023).
- [35] Z. Waks, O. Weissbrod, B. Carmeli, R. Norel, F. Utro, and Y. Goldschmidt, "Driver gene classification reveals a substantial overrepresentation of tumor suppressors among very large chromatin-regulating proteins," *Scientific Reports* 2016 6:1, vol. 6, no. 1, pp. 1-12, Dec. 2016, doi: 10.1038/srep38988.
- [36] A. A. Shah and Y. D. Khan, "Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification," *Scientific Reports* 2020 10:1, vol. 10, no. 1, pp. 1-10, Oct. 2020, doi: 10.1038/s41598-020-73107-y.
- [37] M. D. Levine, "Feature Extraction: A Survey," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1391-1407, 1969, doi: 10.1109/PROC.1969.7277.
- [38] S. J. Malebary and Y. D. Khan, "Evaluating machine learning methodologies for identification of cancer driver genes," *Scientific Reports* 2021 11:1, vol. 11, no. 1, pp. 1-13, Jun. 2021, doi: 10.1038/s41598-021-91656-8.
- [39] M. U. Sohail, J. Shabbir, and F. Sohail, "Imputation of missing values by using raw moments," *Statistics in Transition New Series*, vol. 20, no. 1, pp. 21-40, 2019, doi: 10.21307/STATTRANS-2019-002.
- [40] A. H. Butt and Y. D. Khan, "CanLect-Pred: A cancer therapeutics tool for prediction of target cancerlectins using experiential annotated proteomic sequences," *IEEE Access*, vol. 8, pp. 9520-9531, 2020, doi: 10.1109/ACCESS.2019.2962002.
- [41] M. Shahid, M. Ilyas, W. Hussain, and Y. D. Khan, "ORI-Deep: improving the accuracy for predicting origin of replication sites by using a blend of features and long short-term memory network," *Brief Bioinform*, vol. 23, no. 2, Mar. 2022, doi: 10.1093/BIB/BBAC001.
- [42] M. A. Akmal, N. Rasool, and Y. D. Khan, "Prediction of N-linked glycosylation sites using position relative features and statistical moments," *PLoS One*, vol. 12, no. 8, p. e0181966, Aug. 2017, doi: 10.1371/JOURNAL.PONE.0181966.
- [43] D. K. Barupal and O. Fiehn, "Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach," *Environ Health Perspect*, vol. 127, no. 9, Sep. 2019, doi: 10.1289/EHP4713.
- [44] R. Vinayakumar, M. Alazab, S. Srinivasan, Q. V. Pham, S. K. Padannayil, and K. Simran, "A Visualized Botnet Detection System Based Deep Learning for the Internet of Things Networks of Smart Cities," *IEEE Trans Ind Appl*, vol. 56, no. 4, pp. 4436-4456, Jul. 2020, doi: 10.1109/TIA.2020.2971952.
- [45] D. Rengasamy, M. Jafari, B. Rothwell, X. Chen, and G. P. Figueredo, "Deep Learning with Dynamically Weighted Loss Function for Sensor-Based Prognostics and Health Management," *Sensors* 2020, Vol. 20, Page 723, vol. 20, no. 3, p. 723, Jan. 2020, doi: 10.3390/S20030723.
- [46] J. Gu et al., "Recent advances in convolutional neural networks," *Pattern Recognit*, vol. 77, pp. 354-377, May 2018, doi: 10.1016/J.PATCOG.2017.10.013.

- [47] G. Lin and W. Shen, "Research on convolutional neural network based on improved Relu piecewise activation function," *Procedia Comput Sci*, vol. 131, pp. 977-984, Jan. 2018, doi: 10.1016/J.PROCS.2018.04.239.
- [48] S. Varsamopoulos, K. Bertels, and C. G. Almudever, "Designing neural network based decoders for surface codes," 2018, Accessed: Jan. 14, 2023. [Online]. Available: <https://www.researchgate.net/publication/329362532>
- [49] B. C. Mateus, M. Mendes, J. T. Farinha, R. Assis, and A. M. Cardoso, "Comparing LSTM and GRU Models to Predict the Condition of a Pulp Paper Press," *Energies* 2021, Vol. 14, Page 6958, vol. 14, no. 21, p. 6958, Oct. 2021, doi: 10.3390/EN14216958.
- [50] Y. Gao and D. Glowacka, "Deep Gate Recurrent Neural Network." *PMLR*, pp. 350-365, Nov. 20, 2016. Accessed: Jan. 14, 2023. [Online]. Available: <https://proceedings.mlr.press/v63/gao30.html>
- [51] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602-610, Jul. 2005, doi: 10.1016/J.NEUNET.2005.06.042.
- [52] M. Basaldella, E. Antolli, G. Serra, and C. Tasso, "Bidirectional LSTM recurrent neural network for keyphrase extraction," *Communications in Computer and Information Science*, vol. 806, pp. 180-187, 2018, doi: 10.1007/978-3-319-73165-0_18/COVER.
- [53] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. de Sousa, "Ensemble approaches for regression," *ACM Computing Surveys (CSUR)*, vol. 45, no. 1, Dec. 2012, doi: 10.1145/2379776.2379786.
- [54] L. Breiman, "Bagging predictors," *Mach Learn*, vol. 24, no. 2, pp. 123-140, 1996, doi: 10.1007/BF00058655/METRICS.
- [55] R. E. Schapire, "The strength of weak learnability," *Machine Learning* 1990 5:2, vol. 5, no. 2, pp. 197-227, Jun. 1990, doi: 10.1007/BF00116037.
- [56] S. F. Stefenon et al., "Hybrid Wavelet Stacking Ensemble Model for Insulators Contamination Forecasting," *IEEE Access*, vol. 9, pp. 66387-66397, 2021, doi: 10.1109/ACCESS.2021.3076410.
- [57] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.
- [58] K. M. Ting and I. H. Witten, "Issues in Stacked Generalization," *Journal of Artificial Intelligence Research*, vol. 10, pp. 271-289, May 1999, doi: 10.1613/JAIR.594.
- [59] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A New classifier ensemble method," *IEEE Trans Pattern Anal Mach Intell*, vol. 28, no. 10, pp. 1619-1630, 2006, doi: 10.1109/TPAMI.2006.211.
- [60] "Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [R package e1071 version 1.7-12]," Oct. 2022, Accessed: Jan. 14, 2023. [Online]. Available: <https://CRAN.R-project.org/package=e1071>

- [61] K. J. van Stralen, V. S. Stel, J. B. Reitsma, F. W. Dekker, C. Zoccali, and K. J. Jager, "Diagnostic methods I: sensitivity, specificity, and other measures of accuracy," *Kidney Int*, vol. 75, no. 12, pp. 1257-1263, Jun. 2009, doi: 10.1038/KI.2009.92.
- [62] R. Trevethan, "Sensitivity, Specificity, and Predictive Values: Foundations, Plabilities, and Pitfalls in Research and Practice," *Front Public Health*, vol. 5, p. 307, Nov. 2017, doi: 10.3389/FPUH.2017.00307/BIBTEX.
- [63] A. A. Shah, H. Malik, A. Mohammad, Y. Khan and A. Alourani, "Machine learning techniques for identification of carcinogenic mutations, which cause breast adenocarcinoma", *Scientific Reports*, vol. 12, no. 1, 2022. Available: 10.1038/s41598-022-15533-8.
- [64] Shah, Asghar Ali, Fahad Alturise, Tamim Alkhalifah, and Yaser Daanial Khan. 2022. "Deep Learning Approaches for Detection of Breast Adenocarcinoma Causing Carcinogenic Mutations" *International Journal of Molecular Sciences* 23, no. 19: 11539. <https://doi.org/10.3390/ijms231911539>.
- [65] Shah AA, Alturise F, Alkhalifah T, Khan YD. Evaluation of deep learning techniques for identification of sarcoma-causing carcinogenic mutations. *DIGITAL HEALTH*. 2022;8. doi:10.1177/20552076221133703.
- [66] K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch Ii, "Data analysis and machine learning tools in MATLAB and Python," *Computational Learning Approaches to Data Analytics in Biomedical Applications*. Academic Press Cambridge, MA, USA, pp. 231-290, 2020.
- [67] A. Kulkarni, D. Chong, and F. A. Batarseh, "Foundations of data imbalance and solutions for a data democracy," in *data democracy*, Elsevier, 2020, pp. 83-106.
- [68] A. A. Shah, F. Alturise, T. Alkhalifah, A. Faisal, and Y. D. Khan, "EDLM: Ensemble deep learning model to detect mutation for the early detection of Cholangiocarcinoma," *Genes*, vol. 14, no. 5, p. 1104, 2023. doi:10.3390/genes14051104.
- [69] Ali Shah, A., Shaker, A.S.A., Jabbar, S. et al. An ensemble-based deep learning model for detection of mutation causing cutaneous melanoma. *Sci Rep* 13, 22251 (2023). <https://doi.org/10.1038/s41598-023-49075-4>.