diagnostics

MDPI

Article

A Framework for Prediction of Oncogenomic Progression Aiding Personalized Treatment of Gastric Cancer

Fahad M. Alotaibi [1] and Yaser Daanial Khan [2,*]

[1] Department of Information System, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2] Department of Computer Science, University of Management and Technology, Lahore 54770, Pakistan
* Correspondence: yaser.khan@umt.edu.pk

Abstract: Mutations in genes can alter their DNA patterns, and by recognizing these mutations, many carcinomas can be diagnosed in the progression stages. The human body contains many hidden and enigmatic features that humankind has not yet fully understood. A total of 7539 neoplasm cases were reported from 1 January 2021 to 31 December 2021. Of these, 3156 were seen in males (41.9%) and 4383 (58.1%) in female patients. Several machine learning and deep learning frameworks are already implemented to detect mutations, but these techniques lack generalized datasets and need to be optimized for better results. Deep learning-based neural networks provide the computational power to calculate the complex structures of gastric carcinoma-driven gene mutations. This study proposes deep learning approaches such as long and short-term memory, gated recurrent units and bi-LSTM to help in identifying the progression of gastric carcinoma in an optimized manner. This study includes 61 carcinogenic driver genes whose mutations can cause gastric cancer. The mutation information was downloaded from intOGen.org and normal gene sequences were downloaded from asia.ensembl.org, as explained in the data collection section. The proposed deep learning models are validated using the self-consistency test (SCT), 10-fold cross-validation test (FCVT), and independent set test (IST); the IST prediction metrics of accuracy, sensitivity, specificity, MCC and AUC of LSTM, Bi-LSTM, and GRU are 97.18%, 98.35%, 96.01%, 0.94, 0.98; 99.46%, 98.93%, 100%, 0.989, 1.00; 99.46%, 98.93%, 100%, 0.989 and 1.00, respectively.

Keywords: long and short-term memory (LSTM); bi-LSTM; gated recurrent units (GRU); next generation sequencing (NGS); gastric carcinoma; deep learning; bioinformatics

Citation: Alotaibi, F.M.; Khan, Y.D. A Framework for Prediction of Oncogenomic Progression Aiding Personalized Treatment of Gastric Cancer. *Diagnostics* **2023**, *13*, 2291. https://doi.org/10.3390/diagnostics13132291

Academic Editor: Dechang Chen

Received: 15 March 2023
Revised: 5 June 2023
Accepted: 13 June 2023
Published: 6 July 2023

1. Introduction

Gastric cancer is a malignant cancerous mutation disease. It is the 4th most common cancer among men. Mutation is one of the leading causes of this cancer. Mutation is a genetic disorder that occurs due to changes in the gene sequence. These changes may include deletion, insertion, updation or replication of the gene bases in the gene sequences. The American Joint Commission on Cancer (TNM) divided cancer into four stages: 0, 1, 2, 3, and unstageable. In Pakistan, 6566 new cases were identified in 2020, and 5692 deaths were reported. According to Shaukat Khanum Memorial Cancer Hospital and Research Center (SKMCHRC) 7539 new cases were reported from 1 January 2021 to 31 December 2021. Of these, 3156 were seen in males (41.9%) and 4383 (58.1%) in female patients [1]. The smallest component of DNA, the gene, is a two-fold helix particle made up of direct arrangements of nucleotide sets [2]. Each nucleotide is made up of sequence of gene bases. Gene mutation is a type of gene alteration in which the structure of a gene cell is altered. These mutations can provide details about the development of cancer [3]. As researchers gain a deeper understanding of these mutations, the reasons for asymmetrical carcinoma cell proliferation are growing in number. Gastric carcinoma can be recognized using a variety of biomarkers. Even in the absence of physical symptoms on the body or other

imaging resources used to detect gastric carcinoma, we can still identify gastric carcinoma by identifying patterns in the TCGA of gene mutations [4]. Different forms of carcinomas can be distinguished by focusing on several types of gene mutation. When a mutation occurs in a person's body, it accelerates the growth of certain tumor cells, which results in an increase in the number of active gastric carcinoma cells in the body. This alters the normal cycle of cell genesis and apoptosis [5]. Such alteration forces the death process to stop while the body is still producing new cells. Therefore, the increase in the number of cells in the body is termed gastric carcinoma.

This research aims to make significant contributions in the field of gastric cancer mutations by addressing the limitations of the most recent innovative work. The following is the arrangement of the information in bullet form:

- Explore the development of a universal and explicit benchmark dataset specifically tailored for gastric cancer mutations to overcome existing limitations.
- Investigate potential handcrafted feature extraction techniques to preserve the dataset's integrity and enhance the accuracy of mutation detection models in gastric cancer.
- Examine the shortcomings of current model evaluation methods for accurately assessing the performance of mutation detection models in gastric cancer.
- Propose the development of more robust and comprehensive evaluation techniques to address the limitations of current model evaluation methods.
- Explore the incorporation of improved feature extraction techniques and advanced evaluation methods to enhance the accuracy in the field of gastric cancer mutations.

The proposed study uses the gene sequence dataset for the identification of gastric carcinoma. The most recent and most generalized dataset, as described in the data collection section, was assembled for this study while keeping these limitations in mind. Furthermore, a total accuracy of 99.46% is achieved by utilizing various deep learning methods. Numerous assessments and validation methods are investigated, including the SCT, IST, and 10FCVT. Multiple statistical tools for model evaluation, such as sensitivity, specificity, AUC, and MCC, are also implemented.

2. Related Works

The ability to quickly identify cancer using machine learning is advancing every day. Many papers and research articles have been published on various platforms utilizing various methodologies. Most of these studies uses MRI images for the detection of the gastric cancer. As shown in Table 1, machine learning techniques have become widely used in recent years to provide timely identification models for efficient decision making [6–20].

Table 1. Previous applied algorithms to investigate gastric carcinoma.

Paper Citation	Algorithm	Accuracy Achieved	Dataset
[6]	Adaptive Neural-Fuzzy Inference System	86.00%	PET-Scan, CT-Scan
[7]	Densely Connected Convolutional Network	96.79%	Endoscopy Images
[8]	Logistic Regression	73.20%	Electronic Health Record
[9]	Naive Bayes	74.90%	Gene Expression Data
[10]	Support Vector Machine	70.00%	miRNA
[11]	Extra Tree Classifier Random Forest Classifier Bagging Classifier HGB Classifier LGBM Classifier Decision Tree Classifier Gradient Boost Classifier	97.27% 95.64% 95.21% 95.29% 92.71% 85.75% 79.54%	Surveillance, Epidemiology and End Results (SEER)

Several studies have explored different machine learning techniques for the detection and classification of gastric cancer. A notable approach is the adaptive neural-fuzzy inference system (ANFIS) [6], which achieved an accuracy of 86.00% using PET-Scan and CT-Scan data. Densely connected convolutional networks (DenseNet) [7] demonstrated promising results, with an accuracy of 96.79% when analyzing endoscopy images. Logistic regression [8] was applied to electronic health records and achieved an accuracy of 73.20%. Naive Bayes [9] was utilized for gene expression data, achieving an accuracy of 74.90%. Support vector machines (SVM) [10] achieved an accuracy of 70.00% when analyzing miRNA data.

Furthermore, ensemble learning methods have been employed to improve accuracy. Extra tree classifier, random forest classifier, bagging classifier, and HGB classifier achieved accuracies of 97.27%, 95.64%, 95.21%, and 95.29%, respectively [11]. These classifiers were employed using data from the Surveillance, Epidemiology, and End Results (SEER) database. Other classification algorithms were also explored. The LightGBM (LGBM) classifier achieved an accuracy of 92.71%, while the decision tree classifier achieved 85.75% accuracy. The gradient boost classifier attained an accuracy of 79.54%. Despite the notable performance of these existing approaches, limitations persist.

The most recent innovative works in the field of gastric cancer mutations have encountered various limitations that necessitate attention. One notable constraint is the absence of a universal and explicit benchmark dataset exclusively focused on gastric cancer mutations. To address this, further efforts should be made to incorporate more handcrafted feature extraction techniques that can effectively preserve the integrity of the actual dataset, thereby enhancing the accuracy of mutation detection models. Moreover, the existing model evaluation methods are deemed insufficient, leaving significant room for improvement in terms of accurately assessing the performance of these models. Thus, it becomes imperative to develop more robust and comprehensive evaluation techniques to effectively measure and enhance accuracy in future research endeavors.

3. Materials and Methods

To detect gastric cancer, this study suggests the use of deep learning techniques such LSTM, Bi-LSTM, and GRU. Figure 1 explains this study's general methodology.

3.1. Benchmark Dataset Collection

The benchmark dataset typically includes tentatively settled unambiguous known patterns. These patterns are additionally utilized for testing purposes. Its purpose is to create a high quality benchmark dataset [21–25] which is different, precise, and applicable. There are 1014 samples with 1948 mutations in the 61 driver genes that are connected to gastric cancer. Normal gene sequences were obtained at https://asia.ensembl.org [26] using Python web scraping code, while mutation information were obtained from http://intogen.org [27], also using Python web scraping code. Then, by putting the mutation information into regular gene sequences, another piece of Python code was built to produce mutated sequences. This gives us mutated sequences, but due to the substantial number of mutations and normal dataset, we used the CDHIT tool with a 100% similarity ratio to remove similar sequences from the normal and mutated sequence dataset, leaving us with an unbalanced dataset. Thus, we balanced the dataset to use it for sample formulation [28], and the process is depicted in Figure 2.

This framework requires intense computational power, and one of the most powerful tools available for the lowest cost was Google Colab pro, with 16 GB of GPU and 26 GB of RAM, which took almost 52 h to complete without any interruption. There were a total of 61 gastric carcinoma active mutated driver genes, which are listed in Table 2. All the related driver gene symbols and numbers of mutations in each gene are listed in Table 2.

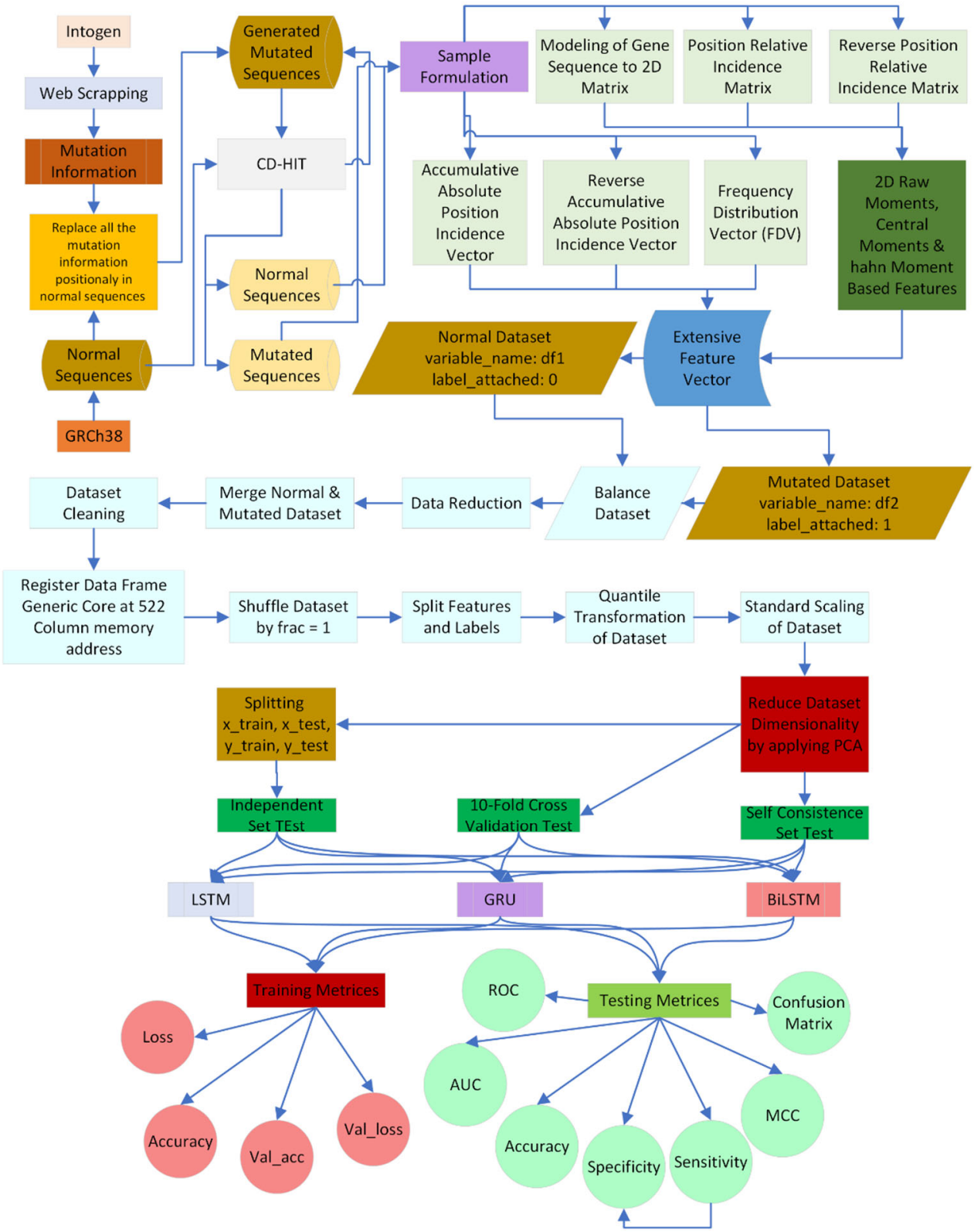

Figure 1. Methodology of the proposed study for identification of mutation to detect stomach carcinoma.

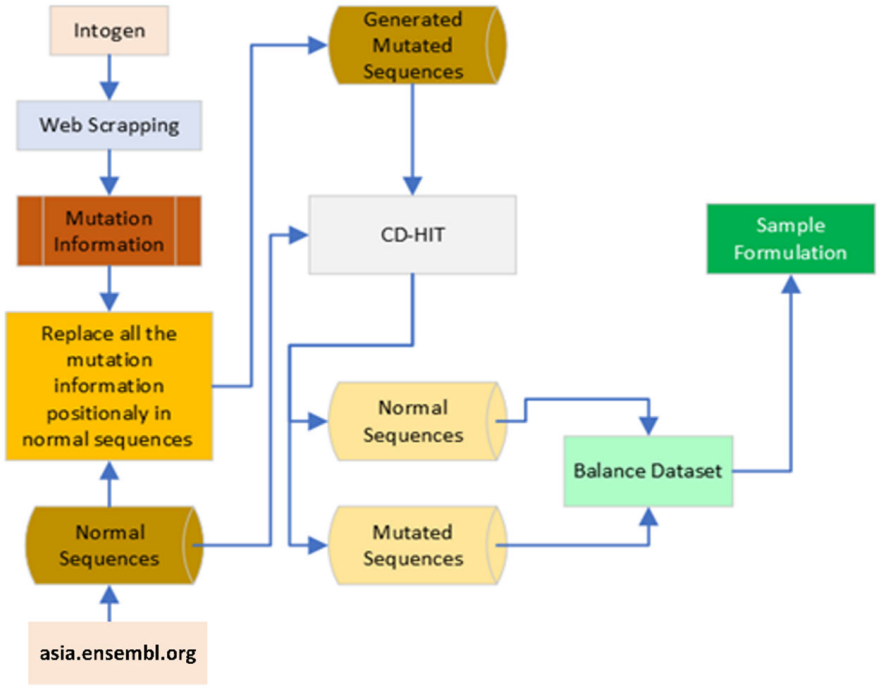

Figure 2. Benchmark dataset collection process.

Table 2. All genes related to stomach cancer with the number of mutations in each gene.

Gene Symbol	No of Mutations	Gene Symbol	No of Mutations	Gene Symbol	No of Mutations
TP53	293	*FBXW7*	16	*ARHGEF12*	13
ARID1A	76	*MAP2K7*	22	*PIK3R1*	5
PIK3CA	75	*SOHLH2*	15	*MYH9*	20
CDH10	52	*NIN*	18	*NTRK3*	17
SMAD4	35	*FAT4*	126	*FAT3*	90
KRAS	37	*PRF1*	15	*BCL9*	14
APC	44	*PRKCB*	14	*ATM*	31
KMT2D	45	*ACVR2A*	24	*KIT*	13
CDH11	33	*RNF43*	16	*CACNA1D*	18
ERBB3	28	*BMPR2*	11	*KDM6A*	11
RHOA	27	*PPP3CA*	9	*CARS*	8
CTNNB1	30	*CASP8*	6	*GRIN2A*	32
LRP1B	169	*TOP2A*	12	*NSD1*	21
ARID2	27	*PRRX1*	9	*FAT1*	31
CDKN2A	18	*ARHGEF10L*	10	*CDK12*	15
BCOR	28	*TET1*	23	*FHIT*	3
ERBB2	26	*RELA*	9	*BCLAF1*	20
DCSTAMP	18	*RB1*	12	*RECQL4*	11
TRIM49C	17	*NRG1*	23	*CLIP1*	10
KMT2C	69	*BMPR1A*	3		
PTEN	20	*SDC4*	5		

3.2. Feature Extraction

Redundancy reduction is helpful for deep learning prediction models which specifically include unsupervised learning. This process helps to support complex data structures, i.e., genes mutation data set. After the successful identification of redundant information, data can be compressed. This can reduce the volume of data without losing any valuable information; only scrappy and messy data, which makes the dataset more complex, are eliminated by this procedure [29]. Extensive feature extraction techniques were developed in this study to prepare the dataset for feeding into the proposed deep learning models, as in Figure 3. Multiple feature extraction techniques were applied in this study, such as reverse accumulative absolute position incidence vector (RAAPIV), accumulative absolute position incidence vector (AAPIV), frequency distribution vector (FDV), modeling of gene sequence to 2D matrix, position relative incidence matrix (PRIM), re-verse position relative incidence matrix (RPRIM), 2D raw moments, central moments, and Hahn moments, as discussed in [30–41], which required extensive research. Equations (2)–(22) describes all the corresponding elements to extract extensive feature vectors based on int64 datatype, which is ideal for LSTM-based architectures. Intelligible and significant information endures import, since the result obtained is a mixture of various unmistakable fair dataset tests. A sizable dataset with a clear description of the malignant growth driver quality successions is put together [42]. As a baseline of genuine malignant growth driver quality sequences, the dataset is required. This work took the benchmark dataset from a very recent version of the interpretation made accessible on the internet, specifically http://intogen.org/ [43]. A sum of 32 malignant growth driver potential genes mutations, i.e., *TP53*, *CDH1*, *SMAD4*, *KRAS*, *APC*, *KMT2D*, *CDH11*, *ERBB3*, *RHOA*, *LRP1B*, *ARID2*, *BCOR*, *ERBB2*, *KMT2C*, *PTEN*, *FBXW7*, *NIN*, *FAT4*, *PRF1*, *PRKCB*, *RNF43*, *BMPR2*, *SDC45*, *ARHGEF12*, *PIK3R1*, *MYH920*, *NTRK317*, *FAT390*, *BCL914*, *ATM31*, *KIT13*, and *CACNA1D*, are associated with gastric carcinoma-causing mutations [24].

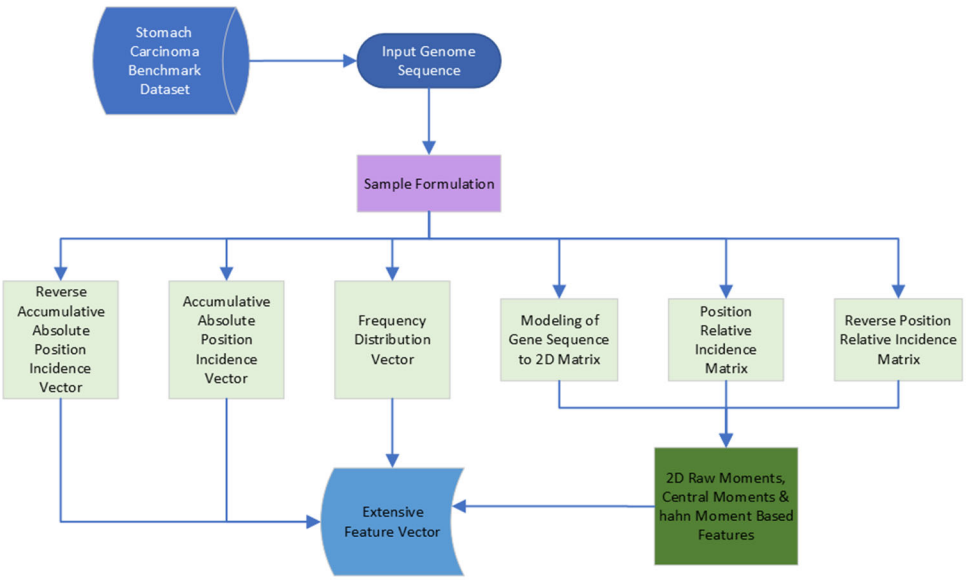

Figure 3. Feature extraction framework.

Thusly, information accumulated in this way is utilized to plan a benchmark dataset The benchmark dataset for gastric carcinoma inside the current review is signified as D, which is characterized as

$$D = D^{+} \, U \, D^{-} \tag{1}$$

The final benchmark dataset included 1948 carcinoma-mutated human gene sequences (D^{+}) and 2000 precisely chosen carcinoma-negative sample genes (D^{-}), acquired from a larger collection of normal genes after careful preprocessing and homology reduction. Gene

sample presentations often employ two diverse types of model development. Most vector formulations use discrete or sequential modelling to represent genomes. The sequential model uses Equation (2) to represent the genome sequence as its nucleotide sequence:

$$S = w_1, w_2, w_3, \ldots w_n \tag{2}$$

where

$$w \in \{A(adenine), C(cytosine), G(guanine), T(thymine)\}$$

where w denotes the nucleotide at any location, and stands for an element contained within the set, with the meaning "member of," [44], the first nucleotide in genome S is represented by w_1, and w_L is the last nucleotide. 'n' represents the total length of the sequence in a genome. The detailing of organic sequencing is one of the most basic issues in computational science. The nucleotide makeup of a genomics sample serves as the discrete model representation in the second model. Equation (3) defines the genome S representation using a discrete model as follows:

$$S = [ds_1\ ds_2\ ds_3\ \ldots\ ds_{20}]^T \tag{3}$$

where the useful component feature $ds_a(a = 1, 2, 3, \ldots 20)$ is represented by the extraction techniques employing pertinent nucleotides in the genome S. These elements are also used in the statistical moment-based feature extraction techniques.

3.2.1. Statistical Moments Calculation

The arrangement of each succession of genes follows some examples. Because of such requirements, each arrangement is portrayed with various measurable boundaries. In past work, factual moments were utilized for highlight extraction [45–47]. To include extraction, crude, focal, and Hahn moments are utilized. The nucleotide component is crucial to the function and makeup of genes. Area and scale variation can be used to extract the component [48]. Crude moments are used to calculate the mean, fluctuation, and imbalance of test appropriation in the dataset in order to address region variation highlights. As mean, difference, and unevenness are assessed using centroid, but focal moments are scaled variably, focal moments are also used for extraction. However, this method is area invariant [29,49]. When measuring measurable limits, Hahn moments are used; however, they come in both area and scale variants [50,51]. In order to evaluate the dataset's mean, variance, and deviation of the probability transmission, Hahn moments are registered using Hahn polynomials. For the aforementioned method, events are recorded in a $n \times n$ two-dimensional grid denoted by A_2D' [42]. For portraying the parts and estimations of Equation (4) and the quantitative depiction of gastric carcinoma driver quality, examples of the benchmarks dataset are used in the real methodology.

This study applied factual moments to change the genomics information to a proper size. Every second portrays some novel data that assigns the idea of information. Examiners and mathematicians have dealt with snapshots of various distributions. Hahn, crude, and focal snapshots of the genomics information are outfitted into the list of capabilities and structures as a striking part of an info vector for the indicator. The region and size of fluctuation integrated into the moments can be used as a device to interpret among practically various groupings. The building of a classifier using the distribution of a labeled dataset also benefits from many moments that define the unbalanced and the average of the information. Researchers have discovered that the design, in addition to the general placement of their bases, affects the characteristics of proteomics and genomics arrangements. From this point forward, only mathematical and statistical models are best suited for outfitting the component vector because they are sensitive to the general positioning of component DNA nucleotides inside genomics successions. It is a basic consideration in forming, yielding, and persevering element sets. Since Hahn moments require two-layered information, the genomics groupings are changed into a two-layered

documentation A' of size $k * k$, which stores a similar amount of data to S, though in a two-layered structure to such an extent that

$$m = \sqrt{n} \tag{4}$$

where 'n' is the sequence length of a sample genome and 'm' represents the $2D$ square matrix dimensions. The A'_{2D} matrix in Equation (5) is formed using the ordering obtained from Equation (4), having '$k \times k$' rows and columns, respectively.

A purpose ω^{29} is a mapping purpose cast-off for matrix transformation of S as A'_{2D}. It uses the component from this matrix A'_{2D}. The raw moments are computed using the values of A'_{2D}. The raw moments of M_{ij}, a $2D$ continuous function with the order $(i + j)$, were computed up to order three, such as $M_{01}, M_{10}, M_{12}, M_{21}, M_{30}$ and M_{03}, and the raw instants are computed as in Equation (6).

$$A'_{2D} = \begin{bmatrix} a_{1\to1} & a_{1\to2}\cdots & a_{1\to j}\cdots & a_{1\to k} \\ a_{2\to1} & a_{2\to1}\cdots & a_{2\to j}\cdots & a_{2\to k} \\ \vdots\ a_{i\to1} & \vdots\ a_{i\to2}\cdots & \vdots\ a_{i\to j}\cdots & \vdots\ a_{i\to k} \\ \vdots\ a_{k\to1} & \vdots\ a_{k\to2}\cdots & \vdots\ a_{k\to j}\cdots & \vdots\ a_{k\to k} \end{bmatrix} \tag{5}$$

$$W_{ij} = \sum_{b=1}^{n}\sum_{q=1}^{n} b^i q^j A'_{2D}(b,q) \tag{6}$$

The order of the instants is indicated by the addition of i and j, that is, $i + j$, which can be less than or equal to three. The above equation's raw moments are computed up to the third order. The origin of the data is used as the starting point from which these raw moments are computed and as the measurement of the separation between the components [45]. The unique characteristics of the raw moments were computed as $W_{00}, W_{01}, W_{10}, W_{11}, W_{02}, W_{20}, W_{12}, W_{21}, W_{30}$ and W_{03}. The centroid of any piece of data is also thought to be its center of gravity. an information point from which the information is evenly spread in all directions. The relationships shown here are those of its weighted average [52]. Using the centroid of the data as their reference point, the central moments unique feature is computed from Equation (7) up to the third order.

$$Q_{ij} = \sum_{b=1}^{n}\sum_{q=1}^{n}(b-\overline{x})^i(q-\overline{y})^j A'_{2D}(b,q) \tag{7}$$

The unique features from central moments, up to the third order, are labeled as $Q_{00}, Q_{10}, Q_{01}, Q_{11}, Q_{02}, Q_{20}, Q_{12}, Q_{21}, Q_{30}$ and Q_{03}. Here, the centroids are calculated as $\overline{x}$ and $\overline{y}$ from Equations (8) and (9):

$$\overline{x} = \frac{M_{10}}{M_{00}} \tag{8}$$

$$\overline{y} = \frac{M_{01}}{M_{00}} \tag{9}$$

Hahn instants can be easily computed for an even-dimensional data body. Reversible possessions of Hahn instants are manifest due to their orthogonality. The square network is utilized as the discrete contribution to figure Hahn moments. Hahn moments assist with depicting the evenness of information and, simultaneously, they are reversible. This essentially implies that these moments can be utilized to reproduce the first information. The reversibility of moments guarantees that the data shortened inside the first arrangement stays in a salvageable shape and is passed forward to the indicator through the relating

highlight vector. Hahn moments are processed utilizing Equation (10), for any integer $r \in [0, P-1]$(P *is a given positive integer*). Hahn instants or order n are computed as

$$h_n^{u,v}(r,P) = (P+v-1)_n (P-1)_n \times \sum_{k=0}^{n} (-1)^k \frac{(-n)_k (-r)_k (2P+u+v-n-1)_k}{(P+v-1)_k (P-1)_k} \cdot \frac{1}{k!} \quad (10)$$

where $(a)_k = a(a+1)\ldots(a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)}$ is the Pochhammer symbol and $u, v (u > -1, v > -1)$ control the shape of polynomials. A square matrix A' is necessary to express the Hahn moments because of its orthogonal features, which necessitate two-dimensional input data. Equation (10) makes use of the Pochhammer documentation, which in turn makes use of the Gamma administrator. Equation (11) explains the Pochhammer symbol:

$$(\Pi)_s = \Pi(\Pi+1)\ldots(\Pi+k-1) \quad (11)$$

The Gamma operator is used to simplify as given in Equation (12):

$$(\Pi)_s = \frac{\Gamma(\Pi+k)}{\Pi(\Pi)} \quad (12)$$

The raw values of Hahn moments given in Equation (13) are often scaled using a weighting function and square norm:

$$\tilde{h_n^{u,v}}(r,P) = h_n^{u,v}(r,P) \sqrt{\frac{\Pi(r)}{k_n^2}}, n = 0, 1, \ldots, P-1 \quad (13)$$

Meanwhile, in Equation (14):

$$\Pi(r) = \frac{\Gamma(x+r+y)\Gamma(y+r+1)(x+y+r+1)_P}{(\mathrm{x}+\mathrm{y}+2\mathrm{r}+1)\mathrm{n}!(\mathrm{P}-\mathrm{r}-1)!} \quad (14)$$

Equation (15) computes the Hahn moments up to third order for the $2D$ discrete data as follows:

$$H_{xy} = \sum_{q=0}^{P-1} \sum_{b=0}^{P-1} A'_{ij} \tilde{h_x^{u,v}}(j,P) \tilde{h_y^{u,v}}(i,P), x, y = 0, 1, \ldots P-1 \quad (15)$$

For every genome sequence, 10 raw, 10 central, and 10 Hahn moments are computed, up to the third order, and are further unified into the collection comprehensive feature vector. These unique features are represented by $H_{00}, H_{01}, H_{10}, H_{11}, H_{12}, H_{21}, H_{20}, H_{02}, H_{30}$ and H_{03}.

3.2.2. Determination of Position Relative Incident Matrix (PRIM)

In next-generation sequencing [53], there are many situations in which the gene arrangements are homologous. This normally happens when a similar predecessor is important for the advancement cycle and more than one grouping is developed from it [54]. In such cases, the exhibition of the classifier is infinitely influenced by utilizing these homologous groupings [55]. Any genome sequence's nucleotide's relative location is regarded as a fundamental pattern that makes use of the physical characteristics of the genome sequence. The genomic sequence is represented by the PRIM in (20×20) order. When managing the results, successful and responsible arrangement resemblance looking is carried out in order to produce correct results. The relative position of each nucleotide in

the given genome sequence is extracted in the form of a matrix, where $Q_{i\to j}$ contains the accumulated worth of jth buildup as for the underlying Equation (16) of the ith buildup.

$$Q_{PRIM} = \begin{bmatrix} Q_{1\to1} & Q_{1\to2}\cdots & Q_{1\to j}\cdots & Q_{1\to20} \\ Q_{2\to1} & Q_{2\to1}\cdots & Q_{2\to j}\cdots & Q_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ Q^{i\to1} & Q^{i\to2}\cdots & Q^{i\to j}\cdots & Q^{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ Q^{k\to1} & Q^{k\to2}\cdots & Q^{k\to j}\cdots & Q^{k\to20} \end{bmatrix} \tag{16}$$

These results represent a replacement of the biological evolutionary process carried out by nucleotides of type "j". A total of 20 native nucleotide occurrences and positional values are shown in alphabetical order. Successful calculations from position relative occurrences in the form of Q_PRIM provide 400 coefficients. The $2D$ Q_PRIM matrix was used to compute 10 Hahn moments, 10 central moments, and 10 raw moments up to 3rd order. Additional 30 distinct features were added before feature extraction.

3.2.3. Determination Reverse Position Relative Incident Matrix (RPRIM)

In AI, exactness and productivity are massively subject to the carefulness and painstakingness of calculations through which the most appropriate provisions in the information are extracted. During the learning stage in AI calculations, learning and transformation of the most implanted obscure patterns in the information are performed to disguise the applicable and relevant elements [47,52,55,56]. RPRIM and PRIM calculations have a similar methodology, yet just RPRIM works with the reverse gene sequence requesting. Processing RPRIM reveals stowed-away patterns that empower the justification of any ambiguities between homologous groupings. It is described by Equation (16). Information is extracted as 400 coefficients for PRIM, which produces a set of 24 elements. Likewise, the above approach is utilized to develop and switch PRIM for a similar succession in a contrary application. The RPRIM is given as Q_{RPRIM}:

$$Q_{RPRIM} = \begin{bmatrix} P_{1\to1} & P_{1\to2}\cdots & P_{1\to j}\cdots & P_{1\to20} \\ P_{2\to1} & P_{2\to1}\cdots & P_{2\to j}\cdots & P_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ P^{i\to1} & P^{i\to2}\cdots & P^{i\to j}\cdots & P^{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ P^{k\to1} & P^{k\to2}\cdots & P^{k\to j}\cdots & P^{k\to20} \end{bmatrix} \tag{17}$$

where $P_{i\to j}$ collected worth of jth buildup concerning the underlaying appearance of the ith buildup utilizing the opposite essential succession. The 2D Q_{RPRIM} matrix was used to compute 10 Hahn moments, 10 central moments, and 10 raw moments up to the third order. The collection of feature extraction was further coordinated to include 30 additional unique features.

3.2.4. Frequency Distribution Vector (FDV)

A frequency distribution vector was created using the distribution of occurrence in each nucleotide of a genomics sequence. Equation (18) defines the frequency distribution vector as follows:

$$\theta = \{\varphi_i, \ldots \varphi_{20}\} \tag{18}$$

Here, the occurrence frequency of ith$(1 \le i \le 20)$ relevant nucleotide is represented as φ_i. However, these techniques are used to reduce information regarding the position importance of nucleotides in a sequence. Additionally, the collection of feature extraction is further coordinated to incorporate 20 features from a frequency distributed vector.

3.2.5. Accumulative Absolute Position Incidence Vector (AAPIV)

Nucleotide distributional information is stored in the frequency distribution vector, but no information on the relative positions of the nucleotides is pertinent. Using AAPIV, 20 relevant nucleotides in a genomic sequence with 20 associated important features might accommodate relative positioning information [48,57]. The collection of feature extraction also coordinates these 20 essential AAPIV traits as shown in Equation (19).

$$AAPIV = \{\beta_i, \ldots \beta_{20}\} \tag{19}$$

Here, β_i is from genome sequence R_x having 'n' total nucleotides, which can be calculated using Equation (20):

$$\beta_i = \sum_{x=1}^{n} R_x \tag{20}$$

3.2.6. Reverse Accumulative Absolute Position Incidence Vector (RAAPIV)

The calculations for RAAPIV and AAPIV follow identical steps; however, only RAAPIV uses the reverse genome sequence ordering. By concealing the deep and hidden patterns of each sample feature, the computing of RAAPIV makes use of reverse relative positioning information [52,58]. The following Equation (21) gives rise to RAAPIV, which produces 20 significant characteristics. These 20 distinct key features from RAAPIV are coordinated with the feature extraction data set.

$$RAAPIV = \{\beta_i, \ldots \beta_{20}\} \tag{21}$$

Here, β_i is from genome sequence R_x having 'n' total nucleotides, which can be calculated using Equation (22):

$$\beta_i = \sum_{x=1}^{n} Reverse(R_x) \tag{22}$$

After features were extracted using the feature extraction approach, 150-D features were created to be used for further processing in the classification algorithm.

3.3. Classification Algorithms

LSTM, GRU, and bi-directional LSTM are deep learning algorithms used in this study. These are also explained in the following subsections.

3.3.1. Long Short-Term Memory (LSTM)

Vanishing gradient problems are solved by applying some specific gates in an RNN, which are built as specified in LSTM and commemorated as $\beth$, explained in Equation (23):

$$\beth = \sigma\left(Wx^{<t>} + Ua^{<t-1>} + b\right) \tag{23}$$

where W, U, b are coefficients specific to the gate and σ is the sigmoid function. The update gate $\beth_u$ defines how much the past should matter, and is used in LSTM. The reset gate $\beth_r$ describes how much previous information should be dropped and is used in LSTM, the forget gate $\beth_f$ defines if a cell should be erased or not and is used in LSTM, and the output gate $\beth_o$ defines how much to reveal of a cell used in LSTM, applying all modification as Equations (24)–(26).

$$c^{<t-1>} = \tanh\left(W_c\left[\beth_r * a^{<t-1>}, \beth_f * a^{<t-1>}, x^{<t>}\right] + b_c\right) \tag{24}$$

$$c^{<t>} = \beth_u * \tilde{c}^{<t>} + \beth_u * c^{<t-1>} + \beth_f * c^{<t-1>} \tag{25}$$

$$a^{<t>} = \beth_o * c^{<t>} \tag{26}$$

The sign $*$ denotes element wise multiplication between two vectors.

The input shape is (64, 1), where 64 represents the number of feature fields, and 1 represents the target field, which can be either positive or negative. During compilation, the loss is calculated using binary cross entropy with the Adam optimizer. The model architecture consists of an input layer followed by an LSTM layer with 128 neurons. After the LSTM layer, there are two dropout layers, a dense layer with 64 nodes, and finally an output layer. Both dropout layers will deactivate 20% of the nodes to avoid overfitting. The output layer has one node with a sigmoid activation function. Figure 4 provides a visual representation of this architecture.

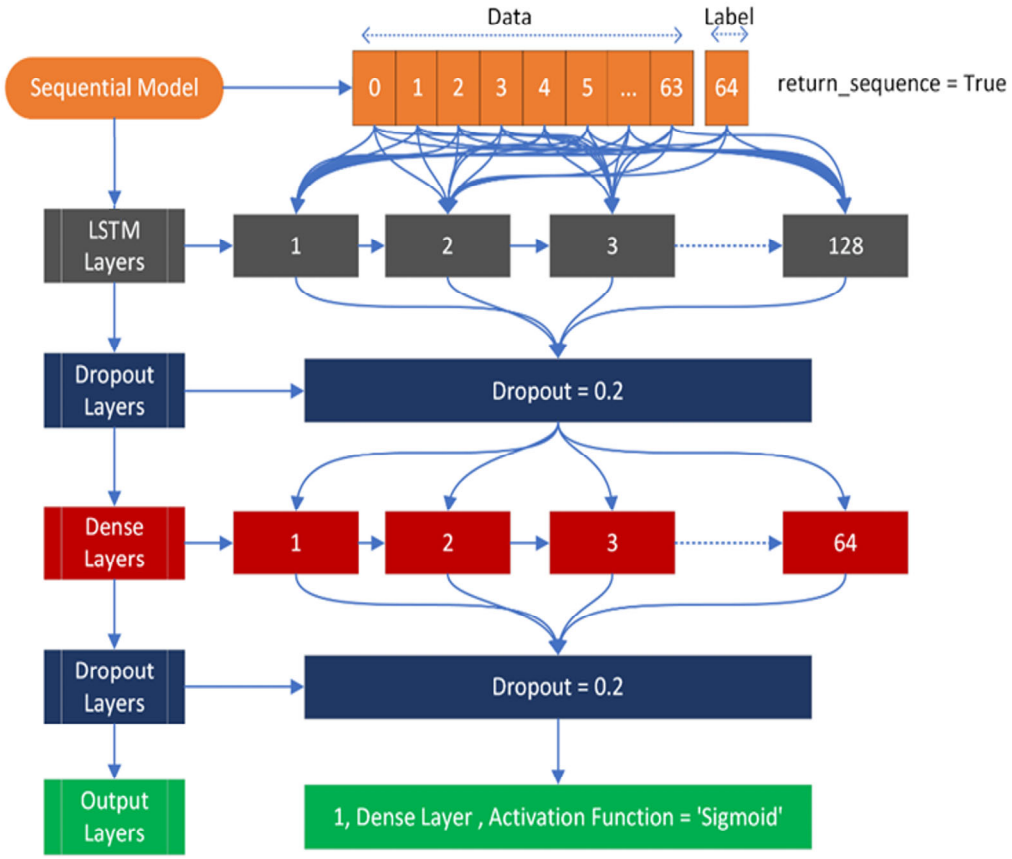

Figure 4. Applied LSTM architecture used to classify mutated and normal gene sequences related to gastric carcinoma.

3.3.2. Gated Recurrent Units (GRU)

The GRU is a more advanced and simple version of the LSTM that was first developed by [59] for application to machine translation. The GRU is based on the LSTM and controls information flow within the unit via update gate $\beth_u$ and Reset gate $\beth_r$ without the use of separate memory. As a result, the GRU can capture the mapping connection between time series data [60], and it also has attractive characteristics such as reduced complexity and an efficient computing procedure, which demonstrates the link between the update and reset gates. The update gate $\beth_u$ defines how much past should matter which is used in the GRU. The reset gate $\beth_r$ describes how much previous information should be dropped and used in the GRU, while the output gate $\beth_o$ defines how much to reveal of a cell used in the GRU, applying all modification as Equations (27)–(29):

$$c^{<t-1>} = \tanh\left(Wc\left[\beth_r * a^{<t-1>}, x^{<t>}\right] + b_c\right) \tag{27}$$

$$c^{<t>} = \beth_u \tilde{c}^{<t>} + (1 - \beth_u) * c^{<t-1>} \tag{28}$$

$$a^{<t>} = c^{<t>} \tag{29}$$

The input shape is (64, 1), where 64 represents the number of feature fields, and 1 represents the target field, which can be either positive or negative. During compilation, the loss is calculated using binary cross entropy with the Adam optimizer. The model architecture includes a GRU layer with 256 nodes, followed by a dropout layer. After the dropout layer, an LSTM layer with 128 nodes is used. The LSTM layer is again followed by a dropout layer, where 20% of the neurons are deactivated to prevent overfitting. Finally, an output layer with a sigmoid activation function is used. Figure 5 provides a visual representation of this architecture.

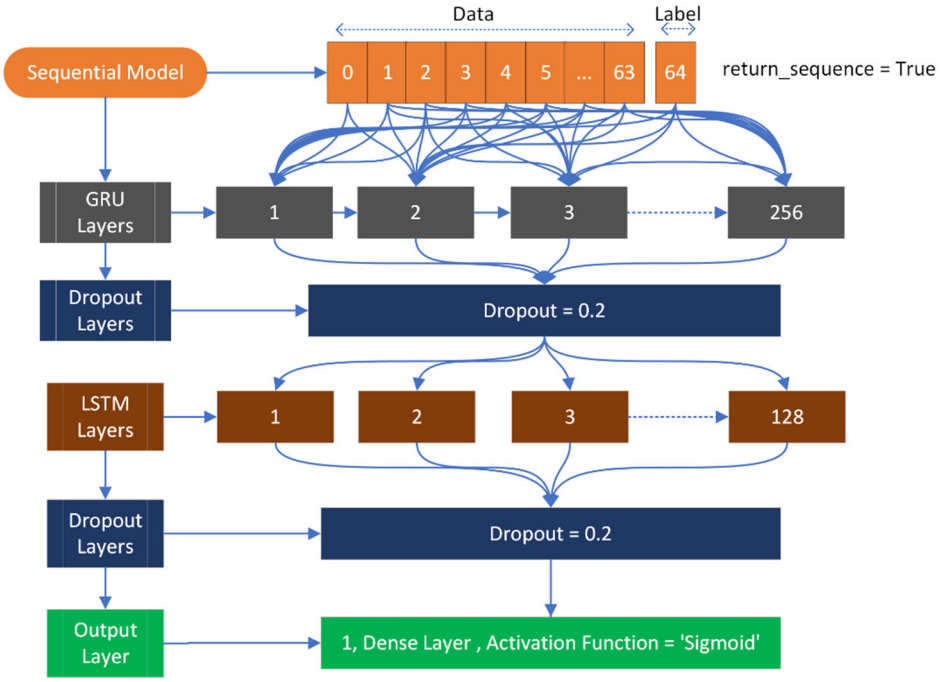

Figure 5. Applied GRU architecture used to classify mutated and normal gene sequences related to gastric carcinoma.

3.3.3. Bidirectional LSTM (Bi-LSTM)

The learning rate scheduling parameter of the LSTM model is tuned using the Adam optimizer. For each variable in the training process, the learning rate is determined adaptively [61]. Adaptive learning rates for various parameters are computed using the first and second moments of gradients. This variant of stochastic gradient descent is known as Adam by the authors.

Nonlinear sigmoidal gates regulate one or more memory cells in a memory block. These gates control whether the model preserves the values at the gates (i.e., the gates evaluate to 1) or discards them (i.e., the gates evaluate to 0). The network computes a mapping sequence to the output $y = (y_1 \ldots\ldots y_T)$ given the input sequence $x = (x_1 \ldots\ldots x_T)$.

Equation (30) can be used to illustrate the fact that information only spreads in the forward direction in LSTM networks, indicating that the state at time t solely depends on the information available before t.

$$\overrightarrow{a^{<t>}} = LSTM\left(x^{<t>}, \overrightarrow{a^{<t+1>}}\right) \tag{30}$$

and when an LSTM backpropagates from the forward direction, then it means direction will be propagated from the last element of the tensor, which can be expresses as Equation (31):

$$\overleftarrow{a^{<t>}} = LSTM\left(x^{<t>}, \overleftarrow{a^{<t+1>}}\right) \tag{31}$$

Finally, the output of the bi-LSTM can be summed as Equation (32) by combining the forward and backward states.

$$a^{<t>} = \left[\overrightarrow{a^{<t>}}, \overleftarrow{a^{<t>}}\right] \tag{32}$$

The input shape is (64, 1), where 64 represents the number of feature fields, and 1 represents the target field, which can be either positive or negative. During compilation, the loss is calculated using binary cross entropy with the Adam optimizer. The model architecture consists of two Bi-LSTM layers, with the first layer having 512 nodes and the second layer having 256 nodes. To avoid overfitting, three dropout layers are used. Additionally, a dense layer with 64 nodes is included. The output layer is a dense layer with a sigmoid activation function to prevent overfitting. Figure 6 provides a visual representation of this architecture.

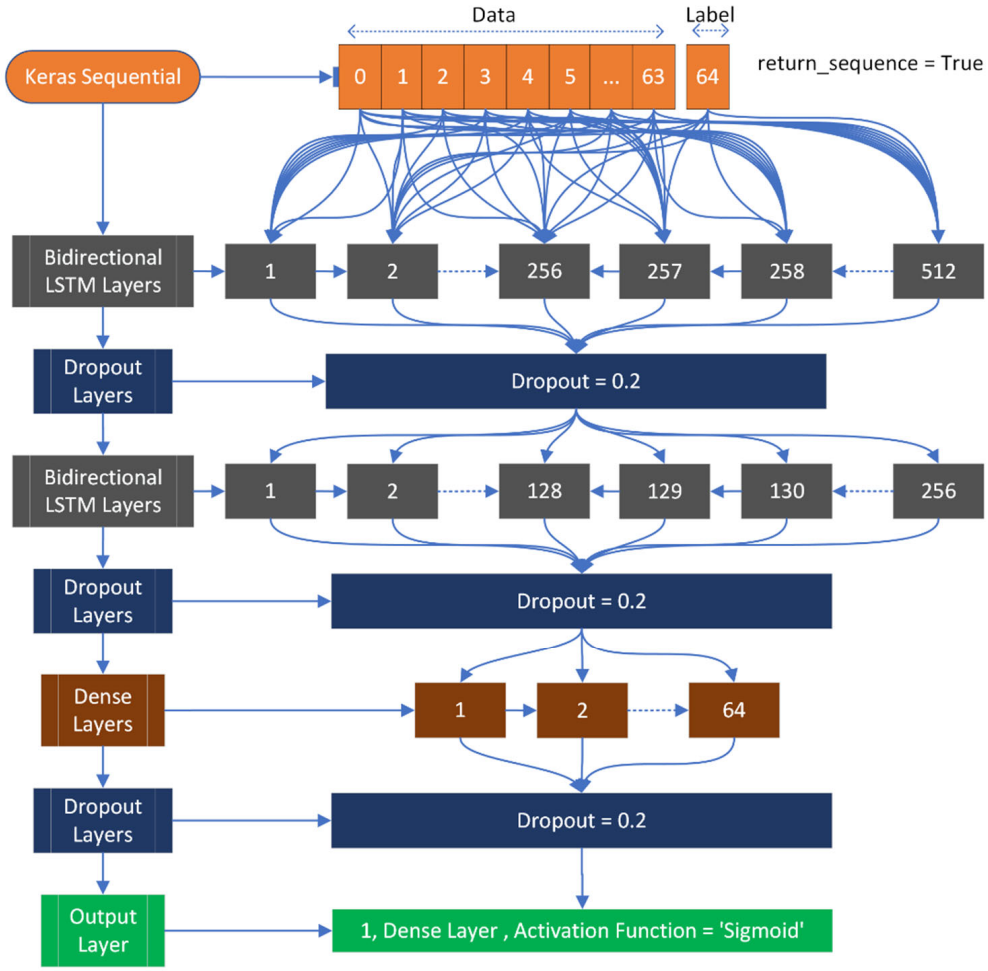

Figure 6. Applied Bi-LSTM architecture used to classify mutated and normal gene sequences related to gastric carcinoma.

4. Results

To measure the performance of the suggested prediction model, it is necessary to compare all the results obtained in this study. A comparison of all the results acquired by this study is shown in Table 3.

Table 3. Comparison of all the obtained results of this study of LSTM, GRU, and bi-directional LSTM.

	Self-Consistency Set Test			**Independent Set Test**			**10-Fold Cross Validation Test**		
Metrics	**LSTM**	**GRU**	**Bi-LSTM**	**LSTM**	**GRU**	**BI-LSTM**	**LSTM**	**GRU**	**Bi-LSTM**
Accuracy (%)	97.18	98.88	98.88	97.18	99.46	99.46	97.30	97.89	97.83
Sensitivity (%)	98.35	100	100	98.35	98.93	98.93	96.10	96.67	96.55
Specificity (%)	96.01	97.77	97.77	96.01	100	100	98.56	99.16	99.16
MCC	0.94	0.977	0.977	0.94	0.989	0.989	0.946	0.978	0.978
AUC	0.98	1.00	1.0	0.98	1.00	1.00	0.99	0.99	0.99

4.1. Self-Consistency Test (SCT)

After complete evaluation, we identified that GRU is best optimized on one notch benchmark dataset. The obtained accuracy, sensitivity, specificity, MCC, AUC of GRU in ISTs are 99.46%, 98.93%, 100%, 0.989, and 1.00, respectively. The obtained result validates the accuracy of the prediction model. This test requires that the indicator is tried with similar examples which were utilized to prepare it. Hereafter, every one of the classifiers prepared on the benchmark dataset is tried. The quantity of tests accurately anticipated by every one of the classifiers is organized to determine the exactness measurements as displayed in Table 3. Thus, the ROC bend shows an examination of precision displayed by every indicator. It is shown that the exhibition of the bi-LSTM indicator is genuinely flourishing when contrasted with GRU and LSTM. Every one of the outcomes yielded by the depicted test is displayed in Table 3. It demonstrates that the predicted rule that was applied during the evaluation was similar to the first computational method that was suggested for the review. The execution of the many different proposed structures that are concerned with this investigation and the evaluation is also demonstrated. Both the training and testing procedures were coordinated with the same dataset in the SCT, because we already know the true positive rate of our benchmark dataset. This test validates the accuracy of training of formulated prediction model. This model does not provide any robust evaluation in the manner of K-fold cross-validation but still has importance in the overall validation process. The results of SCT are given in Table 3. It can be observed that LSTM, Bi-LSTM, and GRU have accuracy values of 97.18%, 98.88%, and 98.88%, respectively. The AUC obtained by LSTM, Bi-LSTM, and GRU is 0.98, 1.00, and 1.00. It validates the correctness of the GRU and Bi-LSTM classifiers. SCT of LSTM model was completed in 63.39 s with a training accuracy of 97.77%. The decision boundary of SCT of LSTM is shown in Figure 7.

There are a total of 100 epochs used to fit the LSTM model in which loss decreased simultaneously from 0.68 to 0.097 in SCT. It shows the compatibility of the dataset with the classifier, and an AUC value of 0.98 shows the optimization of this algorithm on one. There are a total of 100 epochs used to fit the GRU model notch benchmark dataset of gastric carcinoma. The decision boundary of SCT of GRU is shown in Figure 8.

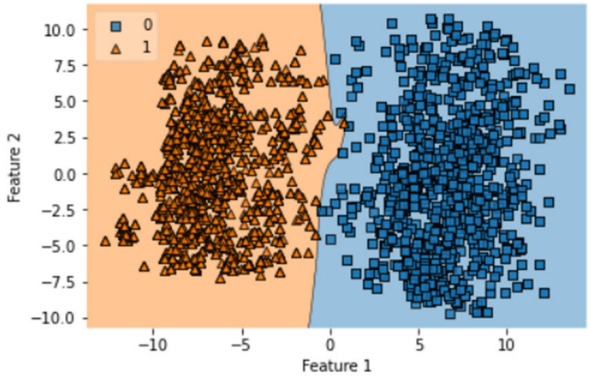

Figure 7. Decision boundary of SCT of LSTM.

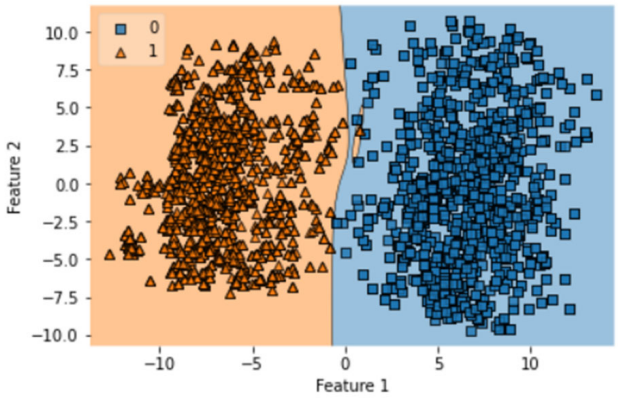

Figure 8. Decision boundary of SCT of GRU.

A total of 100 epochs were used to fit the model for SCT of GRU feature extracted dataset in which loss decreased simultaneously. Moreover, accuracy matrices also increased exponentially, i.e., 53.50 to 100. This behavior shows the exactness of the classifiers with the one-notch benchmark dataset. A total of 100 epochs were used to fit the model with SCT of Bi-LSTM on feature extracted dataset in which loss decreased simultaneously. Moreover, accuracy matrices also increased exponentially, i.e., 92.50 to 100. The decision boundary of SCT of Bi-LSTM is shown in Figure 9. The combined ROC curve of LSTM, GRU and Bi-LSTM is shown in Figure 10. In Figure 10, the green ROC curve illustrates the performance of GRU, while the orange ROC curve represents the performance of Bi-LSTM. The blue dashed line serves as the baseline in the ROC curve, indicating the performance of a random classifier or a model with no discrimination capability.

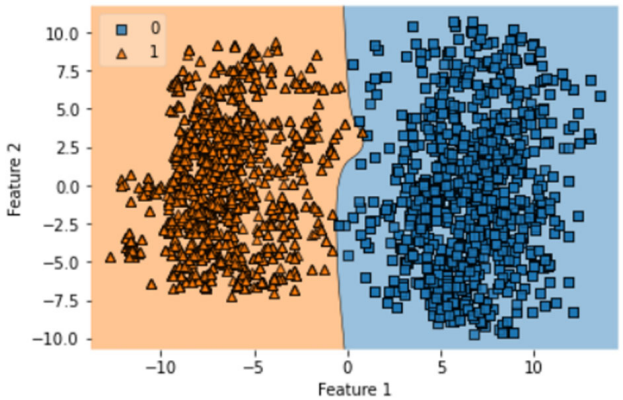

Figure 9. Decision boundary of SCT of Bi-LSTM.

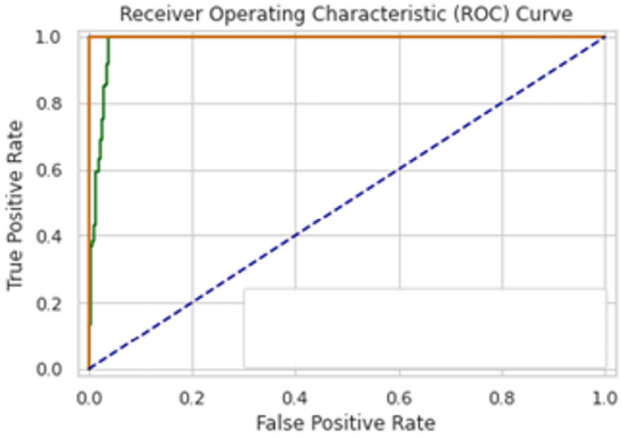

Figure 10. Combined ROC Curve for self-consistency set test of LSTM, GRU, Bi-LSTM.

4.2. Independent Set Test (IST)

A total of 100 epochs were used to fit the model with IST of LSTM on feature extracted dataset in which loss decreased. Moreover, accuracy matrices also increased to 97.77%. The decision boundary of IST of LSTM is shown in Figure 11.

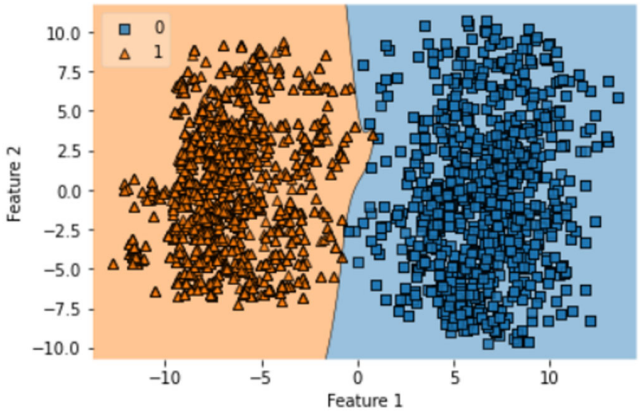

Figure 11. Decision boundary of IST of LSTM.

A total of 100 epochs were used to fit the model for IST of GRU on feature extracted dataset in which loss decreased. Moreover, accuracy matrices also increased from 55.67 to 100. The decision boundary of IST of GRU is shown in Figure 12.

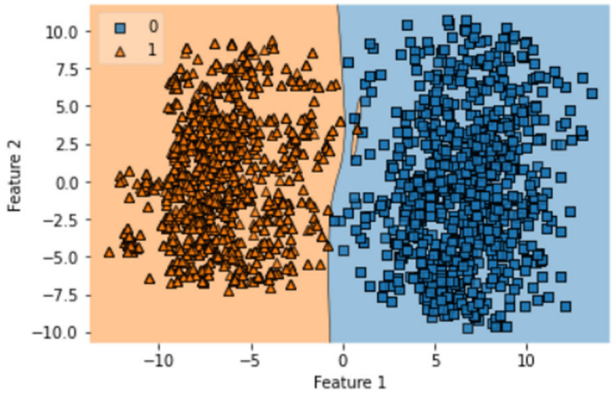

Figure 12. Decision boundary of IST of GRU.

A total of 100 epochs were used to fit the model for IST of Bi-LSTM on feature extracted dataset in which loss decreased. Moreover, accuracy matrices also increased from 99 to

99.82. The decision boundary of IST of Bi-LSTM is shown in Figure 13. The Combined ROC curve is shown in Figure 14. In Figure 14, the green ROC curve illustrates the performance of GRU, while the orange ROC curve represents the performance of Bi-LSTM. The blue dashed line serves as the baseline in the ROC curve, indicating the performance of a random classifier or a model with no discrimination capability.

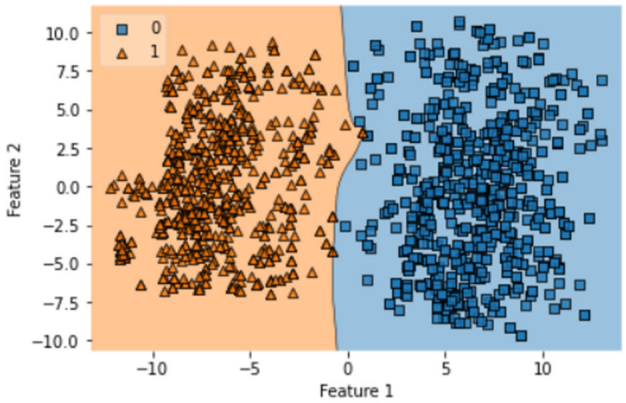

Figure 13. Decision boundary of IST of Bi-LSTM.

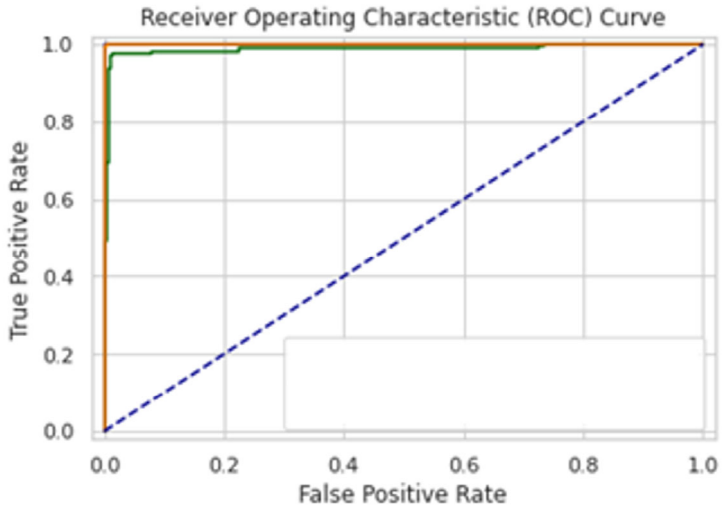

Figure 14. Combine ROC Curve FOR Self-Consistence Set LSTM, GRU, Bi-LSTM.

4.3. 10-Fold Cross-Validation Test (FCVT)

The 10-FCVT sampling test uses a limited number of data samples to validate the formulated prediction model. It has a single parameter, k, which defines how the data sample should be divided. K can be any numeric value; we use k = 10, which folds the overall learning into 10 folds. This it is the best method of validation that predicts true positives. In every fold, a random subset of data is selected for validation from the entire dataset, and accuracy, sensitivity, specificity, and MCC are measured with the mean average value of all fold's results. Detailed results of the 10-FCVT are given in Table 3. It can be observed that the LSTM, Bi-LSTM, and GRU have accuracy values of 97.30%, 97.89% and 97.83%, respectively. The Mean ROC (MROC) values of the LSTM, Bi-LSTM, and GRU are 0.99, 0.99 and 0.99, and given in Figures 15–17, respectively. In Figures 15–17, the green ROC curve illustrates the performance of GRU, while the blue ROC curve represents the performance of LSTM. The blue dashed line serves as the baseline in the ROC curve, indicating the performance of a random classifier or a model with no discrimination capability.

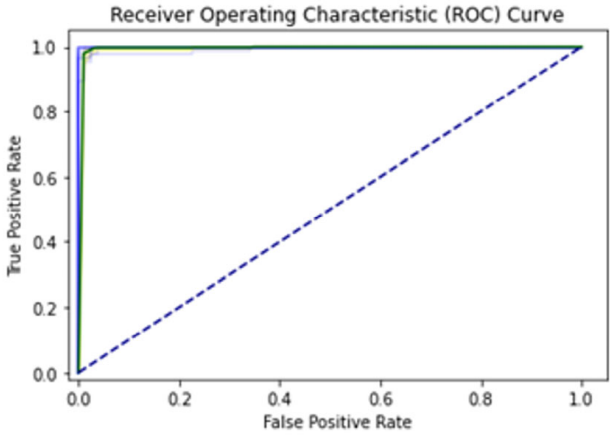

Figure 15. ROC for 10-fold cross validation test of LSTM.

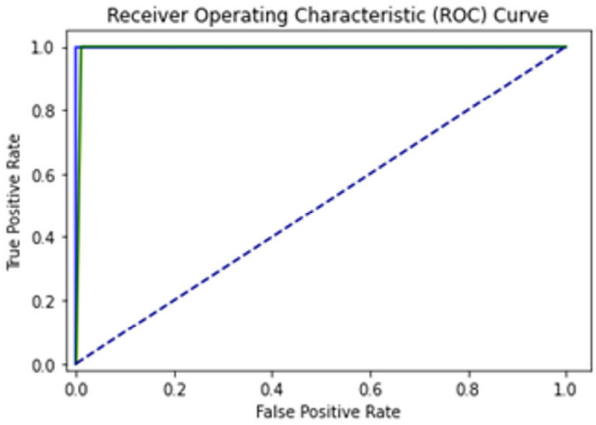

Figure 16. ROC for 10-fold cross validation test GRU.

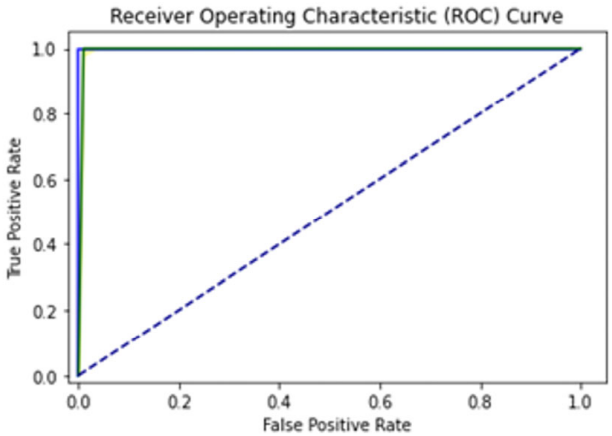

Figure 17. ROC for 10-fold cross validation test Bi-LSTM.

4.4. Comparison with Previous Studies

The independent set test results of LSTM, GRU and Bi-directional LSTM are compared with previous studies in Table 4.

In this study, three different deep learning models were developed, namely LSTM, GRU and Bi-LSTM, and achieved peak accuracies of 97.18, 99.46 and 99.46, respectively. It is clear in Table 4 that this study produced better results than the previous results.

Table 4. Comparison of the current studies with the previous studies.

Current Study		Previous Studies	
Algorithms	**Accuracies Obtained**	**Algorithms**	**Accuracies Obtained**
LSTM	97.18	Adaptive Neural-Fuzzy Inference System [6]	86.00%
GRU	99.46	Densely Connected Convolutional Network [7]	96.79%
Bi-LSTM	99.46	Logistic Regression [8]	73.20%
		Naive Bayes [9]	74.90%
		Support Vector Machine [10]	70.00%
		Random Forest Classifier [10]	95.64%
		LGBM Classifier [11]	92.71%
		Decision Tree Classifier [11]	85.75%
		Gradient Boost Classifier [11]	79.54%

4.5. Complexity Study

A complexity study was conducted to evaluate the impact of incorporated feature extraction techniques developed in this study. Table 5 presents a comparison of the results obtained using feature extraction techniques developed in this study versus the results obtained without utilizing feature extraction techniques developed in this study.

Table 5. A complexity study to assess the contribution of feature extraction techniques developed in this study.

Obtained Results Using Feature Extraction Techniques Developed in This Study		Obtained Results without Using Feature Extraction Techniques Developed in This Study	
Algorithms	**Accuracies Obtained**	**Algorithms**	**Accuracies Obtained**
LSTM	97.18	LSTM	90.42
GRU	99.46	GRU	91.55
Bi-LSTM	99.46	Bi-LSTM	92.67

5. Analysis and Discussion

For the identification and detection of gastric cancer, several biological and computational studies have been conducted. Most researchers used sparse datasets from a small number of hospitals or institutions in previous research, applying machine learning algorithms for detection with lower accuracy and fewer assessment matrices. The most recent generalized huge dataset was employed in deep learning, which included LSTM, BI-LSTM, and GRU for the detection of gastric cancer. The collection includes 1948 mutations in 1014 samples from 61 driver genes related to gastric cancer. The most recent and generalized dataset for the normal and mutant gene sequences of gastric cancer is utilized in this study. Other sorts of mutations are also the subject of a study comparable to this one [62,63], and certain testing methods are also discussed in [64,65]. SCT, IST and 10-FCVT are three separate testing procedures that are applied to the dataset accordingly. It can be inferred from the outcomes of the testing procedures indicated above that the suggested models are most suited to attaining high accuracy for cancer prediction. The entire dataset was used for both the training and testing rounds of the SCT. The results are displayed in Table 3. A total of 80% of the dataset was utilized for training and 20% was used for testing in the IST. The outcomes of ensemble learning utilizing an IST are displayed in Table 3. Ten equal folds were produced from the entire dataset for the 10-FCVT. The proposed deep learning models underwent repeated training on 9-folds and testing on 10-fold. For testing and training, the complete set of data is used. For improved learning, scrambled data are presented each

time, and then the average is determined. The best accuracies were produced by GRU, such as 98.88%, 99.46%, and 97.89% in SCT, IST and 10-FCVT, respectively. Multiple statistical tools for mode evaluation are used in this study. Sensitivity, specificity, MCC, and AUC obtained through GRU in independent tests were 98.93%, 100%, 0.989 and 1.00.

6. Conclusions

This study proposes a framework for identifying the progression of gastric carcinoma by analyzing gene mutations using deep learning-based neural networks. The framework utilizes three RNN variant classifiers: Bi-LSTM, GRU, and LSTM, which were trained on a feature extracted benchmark dataset consisting of 522 fields with labels of either 0 or 1.

The performance and efficiency of the defined models were analyzed using various evaluation metrics, including accuracy, sensitivity, specificity, MCC, and AUC. The results, as presented in Table 3, show that all three models (LSTM, Bi-LSTM, and GRU) achieved high prediction accuracy across different evaluation methodologies (SCT, IST, and 10-FCVT). The metrics demonstrate the models' ability to accurately identify gastric carcinoma progression, with values ranging from 96.01% to 100% for accuracy, 96.10% to 99.46% for sensitivity, 96.55% to 100% for specificity, 0.94 to 0.989 for MCC, and 0.977 to 1.00 for AUC.

These results highlight the potential of deep learning approaches, specifically the proposed framework using RNN variants, in identifying and predicting the progression of gastric carcinoma. However, it is important to note that further optimization and refinement of these strategies and frameworks are necessary to improve their overall performance and achieve even better results.

This study focused on the important task of mutation detection for early detection of gastric cancer. Our work has aimed to contribute to society by addressing the limitations in the current approaches and proposing novel methodologies to enhance the accuracy and efficiency of mutation detection in gastric cancer. The potential impact of our research is significant. Early detection of gastric cancer can significantly improve patient outcomes and survival rates. By accurately identifying and characterizing genetic mutations associated with gastric cancer, our work can contribute to the development of more precise diagnostic tools and targeted therapies. This can lead to earlier intervention, personalized treatment approaches, and improved prognoses for patients. However, it is important to acknowledge the limitations of our work. We recognize that our proposed methodologies may still have room for improvement and require further validation on larger and diverse datasets. Additionally, the complexity of genetic mutations and the heterogeneity of gastric cancer present ongoing challenges in achieving perfect accuracy in mutation detection.

Future Work

Future work in this field should focus on dataset expansion to include more diverse samples, improving the feature extraction process through investigation and optimization, exploring alternative deep learning architectures for model refinement, considering additional evaluation metrics for a comprehensive assessment of model performance, conducting external validation on independent datasets, and collaborating with medical professionals for clinical translation. These efforts will contribute to the continuous improvement of computational methods for identifying and predicting the progression of gastric carcinoma, leading to earlier detection, enhanced treatment strategies, and improved patient outcomes.

Author Contributions: Conceptualization, F.M.A. and Y.D.K.; methodology, F.M.A. and Y.D.K.; validation, F.M.A. and Y.D.K.; resources, F.M.A.; data curation, F.M.A. and Y.D.K.; writing—original draft preparation, F.M.A. and Y.D.K.; writing—review and editing, Y.D.K.; visualization, F.M.A.; supervision, Y.D.K.; project administration, Y.D.K. and F.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by Institutional Fund Projects under grant no. (IFPIP: 1187-830-1443). The authors gratefully acknowledge technical and financial support provided by the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This research work was funded by Institutional Fund Projects under grant no. (IFPIP: 1187-830-1443). The authors gratefully acknowledge technical and financial support provided by the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pisani, P.; Bray, F.; Parkin, D. Estimates of the world-wide prevalence of cancer for 25 sites in the adult population. *Int. J. Cancer* **2001**, *97*, 72–81. [CrossRef] [PubMed]
2. Arshad, A.; Khan, Y. DNA Computing A Survey. In Proceedings of the 2019 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 1–2 November 2019.
3. Loewe, L.; Hill, W. The population genetics of mutations: Good, bad and indifferent. *Philos. Trans. R. Soc. B Biol. Sci.* **2010**, *365*, 1153–1167. [CrossRef] [PubMed]
4. Pareek, C.; Smoczynski, R.; Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet.* **2011**, *52*, 413–435. [CrossRef]
5. Kourou, K.; Exarchos, T.; Exarchos, K.; Karamouzis, M.; Fotiadis, D. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef] [PubMed]
6. Mahdi, A.; Omid, H.; Kaveh, S.; Hamid, R.; Alireza, K.; Alireza, T. Detection of small bowel tumor in wireless capsule endoscopyimages using an adaptive neuro-fuzzy inference system. *J. Biomed. Res.* **2017**, *31*, 419. [CrossRef]
7. Sun, M.; Liang, K.; Zhang, W.; Chang, Q.; Zhou, X. Non-Local Attention and Densely-Connected Convolutional Neural Networks for Malignancy Suspiciousness Classification of Gastric Ulcer. *IEEE Access* **2020**, *8*, 15812–15822. [CrossRef]
8. Huang, R.; Kwon, N.; Tomizawa, Y.; Choi, A.; Hernandez-Boussard, T.; Hwang, J. A Comparison of Logistic Regression against Machine Learning Algorithms for Gastric Cancer Risk Prediction Within Real-World Clinical Data Streams. *JCO Clin. Cancer Inform.* **2022**, *6*, e2200039. [CrossRef]
9. Yang, Y.; Zheng, Y.; Zhang, H.; Miao, Y.; Wu, G.; Zhou, L.; Wang, H.; Ji, R.; Guo, Q.; Chen, Z.; et al. An Immune-Related Gene Panel for Preoperative Lymph Node Status Evaluation in Advanced Gastric Cancer. *BioMed Res. Int.* **2020**, *2020*, 8450656. [CrossRef]
10. Wang, Y.; Wei, Y.; Fan, X.; Xu, F.; Dong, Z.; Cheng, S.; Zhang, J. Construction of a miRNA Signature Using Support Vector Machine to Identify Microsatellite Instability Status and Prognosis in Gastric Cancer. *J. Oncol.* **2022**, *2022*, 6586354. [CrossRef]
11. Polash, M.; Hossen, S.; Sarker, R.; Bhuiyan, M.; Taher, A. Functionality Testing of Machine Learning Algorithms to Anticipate Life Expectancy of Stomach Cancer Patients. In Proceedings of the 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), Gazipur, Bangladesh, 24–26 February 2022.
12. Shah, M.A.; Ud Din, S.; Shah, A.A. Analysis of machine learning techniques for detection framework for DNA repair genes to help diagnose cancer: A systematic literature review. In Proceedings of the 2021 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 9–10 November 2021.
13. Shah, A.A.; Ehsan, M.K.; Sohail, A.; Ilyas, S. Analysis of machine learning techniques for identification of post translation modification in Protein sequencing: A review. In Proceedings of the 2021 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 9–10 November 2021.
14. Ud Din, S.; Shah, M.A.; Shah, A.A. Analysis of machine learning techniques for detection of tumor suppressor genes for early detection of cancer: A systematic literature review. In Proceedings of the 2021 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 9–10 November 2021.
15. Butt, A.H.; Khan, Y.D. Canlect-pred: A cancer therapeutics tool for prediction of Target Cancerlectins using experiential annotated proteomic sequences. *IEEE Access* **2020**, *8*, 9520–9531. [CrossRef]
16. Hussain, W.; Rasool, N.; Khan, Y.D. Insights into machine learning-based approaches for virtual screening in drug discovery: Existing strategies and streamlining through FP-Cadd. *Curr. Drug Discov. Technol.* **2021**, *18*, 463–472. [CrossRef]
17. Khan, Y.D.; Alzahrani, E.; Alghamdi, W.; Ullah, M.Z. Sequence-based identification of allergen proteins developed by integration of pseaac and statistical moments via 5-step rule. *Curr. Bioinform.* **2020**, *15*, 1046–1055. [CrossRef]
18. Naseer, S. NPALMITOYLDEEP-PSEAAC: A predictor of N-palmitoylation sites in proteins using deep representations of proteins and PSEAAC via modified 5-Steps Rule. *Curr. Bioinform.* **2021**, *16*, 294–305. [CrossRef]

19. Naseer, S.; Ali, R.F.; Khan, Y.D.; Dominic, P.D. IGluK-deep: Computational identification of lysine glutarylation sites using deep neural networks with general pseudo amino acid compositions. *J. Biomol. Struct. Dyn.* **2021**, *40*, 11691–11704. [CrossRef] [PubMed]
20. Bashashati, A.; Haffari, G.; Ding, J.; Ha, G.; Lui, K.; Rosner, J.; Huntsman, D.; Caldas, C.; Aparicio, S.; Shah, S. DriverNet: Uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* **2012**, *13*, R124. [CrossRef] [PubMed]
21. Feng, P.; Ding, H.; Yang, H.; Chen, W.; Lin, H.; Chou, K. iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC. *Mol. Ther. Nucleic Acids* **2017**, *7*, 155–163. [CrossRef]
22. Stenson, P.; Mort, M.; Ball, E.; Evans, K.; Hayden, M.; Heywood, S.; Hussain, M.; Phillips, A.; Cooper, D. The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **2017**, *136*, 665–677. [CrossRef]
23. Wu, Y.; Grabsch, H.; Ivanova, T.; Tan, I.; Murray, J.; Ooi, C.; Wright, A.; West, N.; Hutchins, G.; Wu, J.; et al. Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer. *Gut* **2012**, *62*, 1100–1111. [CrossRef]
24. Martínez-Jiménez, F.; Muiños, F.; Sentís, I.; Deu-Pons, J.; Reyes-Salazar, I.; Arnedo-Pac, C.; Mularoni, L.; Pich, O.; Bonet, J.; Kranas, H.; et al. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **2020**, *20*, 555–572. [CrossRef] [PubMed]
25. Zhang, J.; Bajari, R.; Andric, D.; Gerthoffert, F.; Lepsa, A.; Nahal-Bose, H.; Stein, L.; Ferretti, V. The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **2019**, *37*, 367–369. [CrossRef]
26. IntOGen—Cancer Mutations Browser. Available online: https://intogen.org/search (accessed on 2 October 2022).
27. Ensembl Genome Browser 107. Available online: http://asia.ensembl.org/index.html (accessed on 2 October 2022).
28. Guo, S.; Deng, E.; Xu, L.; Ding, H.; Lin, H.; Chen, W.; Chou, K. iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* **2014**, *30*, 1522–1529. [CrossRef]
29. Butt, A.; Rasool, N.; Khan, Y.A. Treatise to Computational Approaches Towards Prediction of Membrane Protein and Its Subtypes. *J. Membr. Biol.* **2016**, *250*, 55–76. [CrossRef] [PubMed]
30. Akcay, M.; Etiz, D.; Celik, O. Prediction of Survival and Recurrence Patterns by Machine Learning in Gastric Cancer Cases Undergoing Radiation Therapy and Chemotherapy. *Adv. Radiat. Oncol.* **2020**, *5*, 1179–1187. [CrossRef]
31. Barukab, O.; Khan, Y.D.; Khan, S.A.; Chou, K.-C. Isulfotyr-PSEAAC: Identify tyrosine sulfation sites by incorporating statistical moments via Chou's 5-steps rule and pseudo components. *Curr. Genom.* **2019**, *20*, 306–320. [CrossRef]
32. Shehryar, S.M.; Shahid, M.A.; Shah, A.A. Mutation detection in genes sequence using machine learning. In Proceedings of the 2021 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 9–10 November 2021.
33. Shah, A.A.; Alturise, F.; Alkhalifah, T.; Khan, Y.D. Evaluation of deep learning techniques for identification of sarcoma-causing carcinogenic mutations. *Digit. Health* **2022**, *8*, 205520762211337. [CrossRef] [PubMed]
34. Hussain, W.; Rasool, N.; Khan, Y.D. A sequence-based predictor of zika virus proteins developed by integration of PSEAAC and statistical moments. *Comb. Chem. High Throughput Screen.* **2020**, *23*, 797–804. [CrossRef] [PubMed]
35. Amanat, S.; Ashraf, A.; Hussain, W.; Rasool, N.; Khan, Y.D. Identification of lysine carboxylation sites in proteins by integrating statistical moments and position relative features via general PSEAAC. *Curr. Bioinform.* **2020**, *15*, 396–407. [CrossRef]
36. Mahmood, M.K.; Ehsan, A.; Khan, Y.D.; Chou, K.-C. Ihyd-LysSite (EPSV): Identifying hydroxylysine sites in protein using statistical formulation by extracting enhanced position and sequence variant feature technique. *Curr. Genom.* **2020**, *21*, 536–545. [CrossRef]
37. Naseer, S.; Hussain, W.; Khan, Y.D.; Rasool, N. IPhosS(deep)-PSEAAC: Identify phosphoserine sites in proteins using deep learning on general pseudo amino acid compositions via modified 5-Steps Rule. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *19*, 1703–1714. [CrossRef]
38. Naseer, S.; Hussain, W.; Khan, Y.D.; Rasool, N. Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations. *Anal. Biochem.* **2021**, *615*, 114069. [CrossRef]
39. Malebary, S.J.; Khan, R.; Khan, Y.D. PROTOPRED: Advancing oncological research through identification of proto-oncogene proteins. *IEEE Access* **2021**, *9*, 68788–68797. [CrossRef]
40. Khan, Y.D.; Khan, N.S.; Naseer, S.; Butt, A.H. Isumok-PSEAAC: Prediction of lysine sumoylation sites using statistical moments and Chou's pseaac. *PeerJ* **2021**, *9*, e11581. [CrossRef]
41. Awais, M.; Hussain, W.; Rasool, N.; Khan, Y.D. ITSP-PSEAAC: Identifying tumor suppressor proteins by using fully connected neural network and PSEAAC. *Curr. Bioinform.* **2021**, *16*, 700–709.
42. Khan, Y.; Rasool, N.; Hussain, W.; Khan, S.; Chou, K. iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. *Anal. Biochem.* **2018**, *550*, 109–116. [CrossRef] [PubMed]
43. Gonzalez-Perez, A.; Perez-Llamas, C.; Deu-Pons, J.; Tamborero, D.; Schroeder, M.; Jene-Sanz, A.; Santos, A.; Lopez-Bigas, N. IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **2013**, *10*, 1081–1082. [CrossRef]
44. Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chen, W.; Chou, K. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* **2019**, *111*, 96–102. [CrossRef]
45. Khan, Y.; Ahmed, F.; Khan, S. Situation recognition using image moments and recurrent neural networks. *Neural Comput. Appl.* **2013**, *24*, 1519–1529. [CrossRef]

46. Khan, Y.; Ahmad, F.; Anwar, W. A neuro-cognitive approach for Iris recognition using back propagation. *World Appl. Sci. J.* **1012**, *16*, 678–685.
47. Akmal, M.; Rasool, N.; Khan, Y. Prediction of N-linked glycosylation sites using position relative features and statistical moments. *PLoS ONE* **2017**, *12*, e0181966. [CrossRef]
48. Ehsan, A.; Mahmood, K.; Khan, Y.; Khan, S.; Chou, K.A. Novel Modeling in Mathematical Biology for Classification of Signal Peptides. *Sci. Rep.* **2018**, *8*, 1039. [CrossRef]
49. Butt, A.; Khan, S.; Jamil, H.; Rasool, N.; Khan, Y. A Prediction Model for Membrane Proteins Using Moments Based Features. *BioMed Res. Int.* **2016**, *2016*, 8370132. [CrossRef]
50. Khan, Y.; Khan, N.; Farooq, S.; Abid, A.; Khan, S.; Ahmad, F.; Mahmood, M. An Efficient Algorithm for Recognition of Human Actions. *Sci. World J.* **2014**, *2014*, 875879. [CrossRef]
51. Khan, Y.; Khan, S.; Ahmad, F.; Islam, S. Iris Recognition Using Image Moments and k-Means Algorithm. *Sci. World J.* **2014**, *2014*, 723595. [CrossRef]
52. Hussain, W.; Khan, Y.; Rasool, N.; Khan, S.; Chou, K. SPrenylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. *J. Theor. Biol.* **2019**, *468*, 1–11. [CrossRef]
53. Grada, A.; Weinbrecht, K. Next-Generation Sequencing: Methodology and Application. *J. Investig. Dermatol.* **2013**, *133*, e11. [CrossRef]
54. Mardis, E. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **2008**, *24*, 133–141. [CrossRef] [PubMed]
55. Awais, M.; Hussain, W.; Khan, Y.; Rasool, N.; Khan, S.; Chou, K. iPhosH-PseAAC: Identify Phosphohistidine Sites in Proteins by Blending Statistical Moments and Position Relative Features According to the Chou's 5-Step Rule and General Pseudo Amino Acid Composition. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 596–610. [CrossRef]
56. Papademetriou, R. Reconstructing with moments. In Proceedings of the 11th IAPR International Conference on Pattern Recognition. Vol. IV. Conference D: Architectures for Vision and Pattern Recognition, The Hague, The Netherlands, 30 August–3 September 1992.
57. Khan, Y.; Jamil, M.; Hussain, W.; Rasool, N.; Khan, S.; Chou, K. pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments. *J. Theor. Biol.* **2019**, *463*, 47–55. [CrossRef] [PubMed]
58. Korthauer, K.; Kendziorski, C. MADGiC: A model-based approach for identifying driver genes in cancer. *Bioinformatics* **2015**, *31*, 1526–1535. [CrossRef]
59. Gruber, N.; Jockisch, A. Are GRU Cells More Specific and LSTM Cells More Sensitive in Motive Classification of Text? *Front. Artif. Intell.* **2020**, *3*, 40. [CrossRef] [PubMed]
60. Reijns, M.; Parry, D.; Williams, T.; Nadeu, F.; Hindshaw, R.; Rios Szwed, D.; Nicholson, M.; Carroll, P.; Boyle, S.; Royo, R.; et al. Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. *Nature* **2022**, *602*, 623–631. [CrossRef]
61. Niu, G.; Wang, X.; Golda, M.; Mastro, S.; Zhang, B. An optimized adaptive PReLU-DBN for rolling element bearing fault diagnosis. *Neurocomputing* **2021**, *445*, 26–34. [CrossRef]
62. Shah, A.; Alturise, F.; Alkhalifah, T.; Khan, Y. Deep Learning Approaches for Detection of Breast Adenocarcinoma Causing Carcinogenic Mutations. *Int. J. Mol. Sci.* **2022**, *23*, 11539. [CrossRef] [PubMed]
63. Shah, A.; Malik, H.; Mohammad, A.; Khan, Y.; Alourani, A. Machine learning techniques for identification of carcinogenic mutations, which cause breast adenocarcinoma. *Sci. Rep.* **2022**, *12*, 11738. [CrossRef] [PubMed]
64. Shah, A.; Khan, Y. Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification. *Sci. Rep.* **2020**, *10*, 16913. [CrossRef] [PubMed]
65. Saeed, S.; Shah, A.; Ehsan, M.; Amirzada, M.; Mahmood, A.; Mezgebo, T. Automated Facial Expression Recognition Framework Using Deep Learning. *J. Healthc. Eng.* **2022**, *2022*, 5707930. [CrossRef]