Enhancing diagnostic accuracy for unidentified primary tumors with remora-optimized novel machine learning approach

Ashok Kumar Singha | Pradeep Kumb | Samir Sapcota

aTeerthanker Mahaveer University, Moradabad, Uttar Pradesh, India, Associate Professor, Department of General Surgery.
bJaipur National University, Jaipur, India, Assistant Professor, Department of Gen Medicine.
cSanskriti University, Mathura, Uttar Pradesh, India, Assistant Professor, Department of Ayurveda.

Abstract Unidentified originating Tumors (UPTs) are a kind of cancer where the originating location of the disease is still unclear, despite cancer cells throughout the body. In other words, even when malignant cells are present, medical professionals are unable to pinpoint the precise organ or tissue where the disease first started. The medical industry offers a big opportunity for Machine Learning (ML) to contribute significantly to illness prediction. One of the primary health challenges that each country faces is the tumor or cancer. In this study, we create a Remora Optimized Gated Recurrent Neural Network (RO-GRNN) to predict UPTs. The proposed approach consists of three stages: pre-processing, classification, and optimization. At first, we classified the UPT using the datasets given in the UCI ML library. The acquired Magnetic Resonance Imaging (MRI)/ Computed Tomography (CT) images will therefore have noise, lowering the efficiency with which classification may be accomplished. Pre-processing techniques like filtering and contrast augmentation could be used to eliminate unwanted noise from the offered images. The Gated Recurrent Neural Network (GRNN) and the Remora Optimization Algorithm (ROA) were both used in the development of the suggested classifier. Multiple optimization procedures, including basic and parametric optimization, take advantage of the ROA inside the GRNN. The proposed approach is employed in Origin Pro, and F1-Measure, recall, precision, and accuracy are some of the performance matrices that are used to evaluate effectiveness. The suggested approach is contrasted with the current ANN, SVM, and LSTM approaches. This research revealed that the proposed method has a 98% accuracy rate for UPT prediction.

Keywords: machine learning, UPT, RO-GRNN, prediction

1. Introduction

Clinical oncology faces major difficulties because of Unidentified Primary Tumors (UPTs). These tumors create diagnostic challenges that impede prompt and efficient treatment choices because they lack a distinct origin. Finding the main location is essential for picking the best treatments, determining the prognosis, and directing patient care. However, since UPTs are so varied and the available diagnostic techniques are so limited in many situations, proper diagnosis is still difficult to achieve (Herruer et al 2020). Despite thorough clinical evaluation, UPT is defined as the accepted metastatic involvement of an organ without a clear primary location. A proper diagnostic strategy produces positive outcomes in these individuals. UPT is regarded as a diagnostic and therapeutic challenge. In 10–30% of individuals with advanced disease, Brain Metastases (BM) is common and has fatal consequences. Brain metastasis often occurs at the end of a patient’s clinical course (Mohamed and Kamel 2021). The accessibility of such information for patients with BM-UPTS is still unknown; however, several papers claimed that the identification rate of original tumors among individuals with extracranial metastases as the initial presentation was 40%.

Despite its crucial clinical implications, metastasis is poorly understood. For instance, it has been reported that metastases may arise from one, many, or clones of the original tumor’s cells, but it is uncertain if these patterns are common across various tumor types. It is also unknown how treatment will affect these patterns and when metastatic seeding will occur (Lin et al 2019). Since metastases are often collected after treatment, with such data, it is hard to distinguish between the factors that cause metastasis and those that are related to treatment because several recent studies have classified genomically unrelated metastatic tumors without the matched primary tumor. However, because of the challenges in collecting matched original tumors and metastases, assessments have been significantly more constrained. There hasn’t been enough research on monoclonal vs. polyclonal planting, when systemic dispersion occurs, or how therapy affects different types of cancer (Hunter et al 2018). As a result, the necessity of diagnostic testing for UPT patients’ therapeutic care...
increased. The main tumor location, tumor size, local-regional metastases, and metastatic sites may all be identified by PET/CT with more accuracy than with contrast-enhanced CT or MRI alone. This makes it easier to choose a treatment that is more suitable and site-specific and to follow up in a way that increases overall survival and improves therapeutic outcomes. Precision head and neck radiation treatment requires laborious tumor target contouring, and radiation oncologists’ approaches to this task vary greatly.

The survival of individuals with head and neck cancer is significantly harmed by contouring errors (Reinert et al 2020). With the increased utilization of proton beam radiotherapy and intensity-modulated radiation therapies, the initial contouring of the Gross Tumor Volume (GTV) has extended dramatically since it is now necessary to imagine data sets that use many modalities or multiple parameters. After an extensive examination of the tumor on diagnostic imaging, the GTV is calculated using the set of CT or MRI information for treatment planning in the traditional way of tumor segmentation (Bakhshayeshkaram et al 2018). Nasopharyngeal carcinoma (NPC) is medically unique from other head and neck malignancies and extraordinarily sensitive to radiation treatment; as a result, radiation therapy is used to treat most of these tumors. The preferred radiation therapy approach for NPC nowadays is intensity-modulated radiation treatments. Because of the following aspects, the NPC may reach nearby cerebral networks and the skull base; GTV contouring for NPC is very labor-intensive and error-prone, but during an MRI, small signal changes often show the extent of participation (Hu et al 2020). To prevent unwanted radiation treatment toxicities, the delineation of the GTV must be precise due to its closeness to the important brain and other organs. Accordingly, the radiation oncologist’s knowledge is essential to designing radiation therapy for NPC. If accessible, deep learning automation of GTV contouring might be a useful situation. The need to increase the diagnostic efficacy of UPTs has gained more attention in recent years. Developing genetic profiling methods, imaging modalities, and computational analysis provides potential new perspectives for improving our comprehension of these puzzling cancers (Fatima et al 2020).

According to the Israel et al (2020), the major uses for these nanoparticles include MRI, magnetic targeting, gene transfer, magnetic hyperthermia for the treatment of tumors, and macrophage polarization for the manipulation of the immune system in the fight against cancer. Because it is more sensitive than MRI, Magnetic Particle Imaging (MPI) has received a lot of interest recently. The most frequent and dangerous primary brain tumors in children and adults exhibit genomic and transcriptional variability. Gonzalez Castro et al (2023) investigated the recognition of the intra-tumoral heterogeneity of primary brain tumors has improved due to the use of these methods to investigate them. They have also learned new information about how these tumors use developmental programs and micro-environmental signaling to promote tumor growth and invasion.

Due to their outstanding performance in image identification tasks, Deep Learning (DL) algorithms are receiving much attention. Medical image properties may be quantitatively evaluated automatically using DL models, leading to improved diagnostic accuracy and more efficiency. According to the Zhou et al (2020), it is feasible to predict clinically undetectable axillary lymph node metastases in individuals with initial breast cancer using US images. Clinically, ultrasound images and biopsies are the best monitoring method for Asian women. Building a model for automated identification, segmentation, and categorization of breast lesions using ultrasound images was the goal of the study (Chiao et al 2019). A method for identifying lesions and differentiating benign from malignant ones was created based on DL and uses mask areas with Convolutional Neural Networks (CNNs). A brain tumor diagnosis approach combining fuzzy logic with edge recognition and the U-NET CNN method of classification was suggested in (Maqsood et al 2021).

The proposed tumor segmentation method uses fuzzy logic image enhancement, edge detection, and classification. Dual-Tree Complex Wavelet Transform (DTCWT) is used at various scale ranges to identify the edge in the input images after contrast improvement, and edge recognition using fuzzier logic has been performed on the original images. The location, shape, and size of a specific form of brain tumor known as a glioma might vary. High-Grade Glioma (HGG) is a more hazardous malignancy than Low-Grade Glioma (LGG). MRI helps assess gliomas in clinical settings because it gives crucial information about the locations of the tumors.

To segment and identify brain tumors, Sharif et al (2020) proposed an active DL-based feature selection technique. The first stage is improving the contrast, which is then given to SbDL for building a saliency map, which is then converted into a binarized form by using straightforward thresholding. Timely and precise identification of a brain tumor is crucial for therapy. Using MRI images, Cinar and Yildirim (2020) aimed to identify a brain tumor. The diagnosing method uses CNN models, one of the deep learning networks. The basis is the Resnet50 architecture, one of the CNN models. The Resnet50 model’s last five layers were deleted, and eight new layers were created in their place. In this research, we use an ML technique called the Remora Optimized Gated Recurrent Neural Network (RO-GRNN) to improve the diagnosis accuracy for UPT.

2. Materials and Methods

This section intends to create the Remora Optimized Gated Recurrent Neural Network (RO-GRNN) for primary tumor prediction. Pre-processing, classification, and optimization are the three stages of the suggested methodology. Figure 1 depicts the whole design of the proposed approach.

https://www.malque.pub/ojs/index.php/msj
2.1. Dataset

The initial step in system processing is data gathering. In this study, the UCI ML primary tumor dataset was used to assess the performance metrics of the two suggested methods. The collection contains 339 instances, 18 characteristics, and one class attribute. The overall number of main tumor classifications is 22. While the "goal" field indicates whether or not the patient has a primary-stage tumor, the other attributes demonstrate the sources of primary tumors. Table 1 describes the key feature qualities. Attributes of a dataset are qualities that are used by systems. In this study, data were preprocessed before the analytical step. One of them is the handling of missing values. As part of this work, we must change a few classed data sets using dummy value meanings that represent 0 and 1. For reliable findings, you need a balanced collection of data. By equalizing both target classes via data balancing, we have increased the validity of the validation.

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Attribute</th>
<th>Description of the Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Demographic data related to patient’s age (&lt;30, 30-59, &gt;=60) in years</td>
</tr>
<tr>
<td>2</td>
<td>Sex</td>
<td>Demographic data related to the Gender of the patient</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male-1, Female-0</td>
</tr>
<tr>
<td>3</td>
<td>degree-of-differ</td>
<td>Multivariate data of 3 degrees well, fairly, poorly</td>
</tr>
<tr>
<td>4</td>
<td>histologic-type</td>
<td>Multivariate data of 3 types</td>
</tr>
<tr>
<td></td>
<td></td>
<td>type1: epidermoid, type 2: adeno, type 3: anaplastic</td>
</tr>
<tr>
<td>5</td>
<td>Brain</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>6</td>
<td>Lung</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>7</td>
<td>Bone</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>8</td>
<td>Liver</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>9</td>
<td>Skin</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>10</td>
<td>Neck</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>11</td>
<td>bone-marrow</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>12</td>
<td>abdominal</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>13</td>
<td>axillary</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>14</td>
<td>mediastinum</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>15</td>
<td>Pleura</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>16</td>
<td>supraclavicular</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
<tr>
<td>17</td>
<td>peritoneum</td>
<td>Yes-1(tumor detected), No-0(not detected)</td>
</tr>
</tbody>
</table>

2.2. Pre-Processing

Preprocessing MRI/CT images include several processes to increase the information required for analysis, enhance image quality, and reduce noise.

Inspect the images to make sure the images are focused appropriately and that the anatomical alignment is constant across all slices. To guarantee constant voxel dimensions, scaling may also be used. Utilize noise reduction methods to
eliminate random noise and raise the signal-to-noise ratio. Many techniques use wavelet demising or spatial filtering, such as Gaussian or median filtering.

The acquired MRI/CT tumor images include undesired sounds, which should lessen the anticipated technique’s ability to function well. The filter reduces the amount of noise in the images. The image has noise in the last few surrounding pixels, and by previously using the average value to alter the pixel quality, pixel worth may range from 0 s to 255 s. The adaptive histogram equalization approach helps to improve contrast after the image has been cleaned of noise. The formulation of the equalization procedure is as follows:

\[
\text{CONTRAST}(j, l) = \text{RANK} \times \text{MAX}_{\text{intensity}(j, l)} \text{ i.e., initially rank } = 0 + 1 \quad (1)
\]

The leading position of the final row across the preceding column it is possible to create the graph in the first position of each line, which contains the new beginning row. The interval, which it continuously identifies as the imaging gray stage, also changes the dispersion of two close gray levels in the current histogram and raises the complexities of the MRI/CT pictures.

2.3. Remora Optimized Gated Recurrent Neural Network (RO-GRNN)

2.3.1. Gated Recurrent Neural Network (GRNN)

We utilize a GRNN framework with Gated Recurrent Units (GRUs) in this study. Recurrent neural networks (RNNs), also known as GRNNs, are a feature of the design of neural networks. Gating mechanisms are included in GRNNs to regulate the flow of information through the network, in contrast to conventional RNNs, which only have one hidden state.

The presence of gating units, often implemented as a mix of sigmoid activation functions and element-wise multiplication operations, is the crucial component of GRNN. These gating components control the information flow between the network’s input, hidden state, and output. GRNNs are designed to deal with the vanishing gradient issue that standard RNNs often run into during training. GRNNs may selectively keep and update data in the hidden state with the help of gating units, allowing them to detect long-term relationships in sequential data. The following are the GRU’s update equations:

\[
y_s = \sigma (X_y u_s + X_y g_{s-1} + a_y) \quad (2)
\]

\[
q_s = \sigma (X_y u_s + X_y g_{s-1} + a_y) \quad (3)
\]

\[
\tilde{g}_s = \text{tanh}(X u_y + X (q_s \odot g_{s-1} + a)) \quad (4)
\]

\[
g_s = (y_s \odot g_{s-1} + (1 - y_s) \odot \tilde{g}_s) \quad (5)
\]

Where \(\sigma(\cdot)\) stands for the elementwise multiplication, the activated sigmoid function \(i\), and \(\odot\) the hyper-tangent activation function. Equations 2-5 are, respectively, the candidate hidden state, update gate, reset gate, and hidden state. \(\{X_y, X_a\} \) are hidden weight matrices that accept input. \(\{V_y, V_a, V\} \) are matrices from hidden to hidden. Bias vectors \(a_y, a_q, a\). The candidate hidden state \(\tilde{g}_s\) is computed using the output of the hidden state \(g_s\), and the reset gate \(q_s\) is updated using the update gate’s result \(y_s\).

One GRU layer makes up our regression model. An 83-point sensor reading series is provided as the input. To predict the tumor concentration, the final hidden state is run through a linear regressor. In this paper, several GRU models are evaluated using layer sizes of 50, 100, and 200 units.

2.3.2. Remora Optimization Algorithm (ROA)

The capacity of the remora to swim alongside whales or other marine creatures to preserve power and avoid predation is well known. The shipping industry often uses tropical waterways. However, like a virus, it also displaces into the host’s water. The remora’s primary supply of nourishment is other fish or invertebrates. It leaves the premises when it locates a region in the ocean with adequate food, ingests it, and then absorbs it back onto the host to continue its voyage to more sponsors and other marine locations. With the number of remora and different dimensions, the ROA algorithm is launched at the present location in the search space. Here is an example of the initial operations phase:

\[
q_1 = (q_{j1}, q_{j2}, \ldots, q_{jC}) \quad (6)
\]

\(J\) is the number of remoras, and \(C\) is the remora’s capacity specified.

This place is designated for the updating procedure with a comparable timeframe since the remora is linked to the swordfish. The formulation of this method is as follows:

\[
q^{s+1}_j = q^{s}_{best} \left[ \text{RAND}(0, 1) \times \left( \frac{q^{s}_{best} + q^{s}_{RAND}}{2} \right) - q^{s}_{RAND} \right] \quad (7)
\]

https://www.malque.pub/ojs/index.php/msj
Where $q_{\text{RAND}}$ is the random position, $S$ is the greatest number of repetitions, while $s$ is the present amount of repetitions being carried out. The elite select the historically wisest course of action, which continues the update process. The pseudocode of ROA is shown in Algorithm 1.

Algorithm 1: Pseudocode of ROA

Set the population and storage location for the random place at the beginning.
Choose the ideal option that is also connected to optimum fitness.
When $S > S_{\text{max}}$
Determine each remora's health benefits parameter
Check to see if any search agents go outside of the search area
Update $\alpha, A, U$
For each remora listed by $I$
If $G(f) = 0$ then
The location of the linked whales is updated using Equation (7)
Else if $G(f) = 1$ then
The associated sailfishes' location is updated using Equation (5)
End if
Calculate (6) the one-step identity.
Calculate the parameter $G(f)$ using Equations (7) and (8) to determine if a host change is necessary.
Equation (12) may be used as the host feeding mode for remora if the host is left unchanged.
End for
End while

The experience of attack is considered while calculating the necessary change of host. The formulation of the experience attack is as follows:

$$q_{\text{att}} = q_f^s + (q_f^s - q_{\text{pre}}) \times \text{randn} \quad (8)$$

Here, $q_{\text{att}}$ is referred to as a tentative step, and $q_{\text{pre}}$ is referred to as the position of the preceding generation.

The following is a description of the status of the formula change based on the standard WOA technique:

$$q_{f+1} = c \times F_b \times \cos(2\pi\alpha) + q_1 \quad (9)$$

$$\alpha = \text{RAND}(0,1) \times (b - 1) + 1 \quad (10)$$

$$b = -\left[1 + \frac{s}{2}\right] \quad (11)$$

$$C = |q_{\text{best}} - q_f| \quad (12)$$

In a higher resolution time, the remora could represent a whale, and its locations would be reduced accordingly. In this case, $e$ stands for the distance between the hunter and the prey while for the random number in the range $[1, 1]$. Furthermore, $[2, 1]$ appears as a random number.

This is an exploiting strategy. At this point, less solution space is available in the main location region. Following is a formulation for the crowd going on the little steps:

$$q_f^s = q_f^s + B \quad (13)$$

$$B = A \times (q_f^s - d \times q_{\text{best}}) \quad (14)$$

$$A = 2 \times Y \times \text{RAND}(0,1) - Y \quad (15)$$

$$Y = 2 \times (1 - \frac{s}{\text{Maximum iteration}}) \quad (16)$$

Here, $D$ represents the location of the remora, while $B$ is a representation of minor motions that could be related to the host volumetric area.

According to the details that have been presented, it would seem that the ROA method is used to optimize the parameters for the GRNN classifier to increase the performance of the classifier when it comes to predicting UPT. During optimization, the GRNN classifier's parameters are evaluated using a fitness assessment to determine which ones should be changed.
3. Results

Performances of the proposed method are assessed using performance matrices, including accuracy, precision, recall, and F1-Measure for the suggested approach, which is implemented in Origin Pro. TP for True positive, TN for True negative, FP for False positive, and FN for False negative. The proposed strategy is evaluated by contrast with well-established methods like Artificial Neural Network (ANN) (Muhammad et al. (2019)), Support Vector Machine (SVM) (Manju et al. (2021)), and Long Short-Term Memory (LSTM) (Amin et al. (2020)), respectively, to ensure its validity.

3.1. Accuracy

Accuracy is a term used to describe how successfully a prediction model recognizes the presence or absence of a tumor in a given dataset when discussing tumor prediction. It is often reported as a percentage and is determined by dividing the total number of instances by the number of cases that were properly predicted (both true positives and true negatives). The accuracy comparison between the suggested technique and current methods is shown in Figure 2 and Table 2. The formula for accuracy is:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{17}
\]

3.2. Precision

Precision is a statistical parameter that assesses the accuracy of the optimistic predictions produced by a predictive model or algorithm in the setting of tumor prediction. Out of all occurrences anticipated as positive (including true positives and false positives), it measures the percentage of accurately predicted positive cases (true positives). The accuracy comparison between the suggested technique and current methods is shown in Figure 3 and Table 3. The formula for precision is:

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{18}
\]
3.3. Recall

The recall is a performance statistic that is often used in the field of tumor prediction to assess how accurate a model or algorithm is. Recall, sometimes referred to as sensitivity or the percentage of instances that the prediction model properly classified as positive (i.e., those in which a person develops a tumor), quantifies the proportion of real positive cases. It offers data on the model’s accuracy in identifying those who are at tumor development risk. The recall comparison between the suggested technique and current methods is shown in Figure 4 and Table 4. The formula for the recall is:

\[ \text{Recall} = \frac{TP}{TP + FN} \]  

(19)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>94.3</td>
</tr>
<tr>
<td>SVM</td>
<td>93.6</td>
</tr>
<tr>
<td>LSTM</td>
<td>80.7</td>
</tr>
<tr>
<td>RO-GRNN [Proposed]</td>
<td>95</td>
</tr>
</tbody>
</table>

3.4. F1-measure

The frequently used measurement for assessing the efficacy of a binary classification model is the F1 measure. It provides a fair assessment of the model’s accuracy by combining recall and precision. The harmonic mean of accuracy and
memory, known as the F1 measure, yields a single score for precision and recall. The F1-measure comparison between the suggested technique and current methods is shown in Figure 5 and Table 5. The formula used to compute it is as follows:

\[
F1 - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]  

(20)

According to the findings above, the suggested approach is 98% accurate. The ANN, SVM, and LSTM also obtained 95.1%, 94.3%, and 93.6%, respectively. The suggested method’s precision score is 97%. The ANN, SVM, and LSTM also obtained 96.2%, 82.5%, and 87.6%, respectively. The suggested method’s recall measurement is 95%. The ANN, SVM, and LSTM obtained 94.3%, 93.6%, and 80.7%, respectively. The suggested method’s F1-measure is 94%. The ANN, SVM, and LSTM also obtained 92%, 87.4%, and 81.9%, respectively. These results demonstrate that RO-GRNN outperforms other presently available methods.

4. Conclusions

Early tumor detection is important for clinical care and cancer monitoring and is a prominent area of current study. It may enhance a patient’s quality of life. The focus of this study, which used ML approaches, was on UPT prediction utilizing a variety of algorithms and a combination of numerous targeted characteristics. Depending on the dataset, a specific ML approach was used, and it was often found that binary classifiers were effective with no more than two categories, classifiers incorporating binary classifiers, called RO-GRNN, were more precise with more classes. It has been compared with traditional methods like ANN, SVM, and LSTM to verify the predicted methodology. The accuracy, precision, recall, and F-Measure of the suggested approach are 98.2%, 97%, 95%, and 94%, respectively. According to the study, the proposed approach produced effective results in terms of statistical measures. Future development and enhancement of the work may automate the detection of initial cancers. When working with big datasets, the suggested approach takes a long time. The development of primary tumor detection will be aided by considering a new, effective strategy to help overcome the disadvantages. This is a solid closing point that we would want to make: the settings of the learning algorithms weren’t optimized to produce the best results with our data. Therefore, there is still an opportunity for improvement.

Ethical considerations

Not applicable.

Declaration of interest

The authors declare no conflicts of interest.
Funding
This research did not receive any financial support.

References


